Casualty Actuarial Society
E-Forum, Spring 2018-Volume 2

The CAS *E-Forum*, Spring 2018-Volume 2

The Spring 2018, Volume 2 edition of the CAS *E-Forum* is a cooperative effort between the CAS *E-Forum* Committee and various CAS committees, task forces, working parties and special interest sections.

This *E-Forum* contains the 2018 CAS Ratemaking Call Papers, the 2018 CAS Climate Change Call Papers and three independent research papers.

CAS *E-Forum*, Spring 2018-Volume 2

Table of Contents

Ratemaking Call Papers

Climate Change Call Papers

Independent Research

E-Forum Committee

For information on submitting a paper to the *E-Forum*, visit http://www.casact.org/pubs/forum/.

PTBA - Risk Selection In Cyber Insurance Underwriting

Ari Chatterjee, ACAS & Dr Raveem Ismail

ABSTRACT

Cyber is an emerging line of insurance, which has demonstrated tremendous growth potential over the next decade. Since it is also an anthropogenic peril, with evolving threat landscape and coverages, it is naturally challenging to underwrite. Here, we propose a new and simple measure, the PTBA (Propensity To Be Attacked). Its key advantages are that it is simple to calculate, and is driven by the interplay between attacker motivation and cybersecurity defence. It produces a single number as an output, and is therefore an ideal *risk score*, a familiar concept in the insurance world (e.g., the terrorism class), and pivotal to quick and practical relative risk appraisal required for underwriting decisions.

Keywords: Cyber, Insurance, Reinsurance, Underwriting, Pricing, Risk, Modelling, Catastrophe.

Table of Contents

1. RESEARCH CONTEXT & OBJECTIVE

Given the paucity of literature on cyber risk insurance, pricing and underwriting, this paper aims to outline a method to underwrite and select risks in a dynamic cyber threat landscape. The traditional methodology of risk classification fails to capture the dynamic nature of the threat landscape and very often, data collected by insurers is insufficient for constructing sophisticated risk classes. We believe the proposed will assist underwriters and actuaries in profitably underwriting cyber insurance.

2. DEFINING PROPENSITY TO BE ATTACKED (PTBA)

The expected income to an attacker from a cyber-attack is the value of each record hacked, plus any other value that might be derived from the target. I.e., if **I** is the *expected income*, then:

$$I = N_{PII}C_{PII} + N_{PHI}C_{PHI} + O\ ,$$

where:

- **N** is the number of records the attacker expects to exfiltrate from a target firm.
- **C** is the expected price per record.
- **O** is *other gains expected by attacker from target* (includes ransom, IP, possible trading insights, possibility of gaining access to larger targets, recognition, etc.).
- **PII** is Personally Identifiable Information (as defined by NIST, e.g., name, date of birth, credit card information, email address, etc.).
- **PHI** is Protected Health Information (as defined by HIPAA, e.g., names, medical records, biometric details, etc.). The estimated relative value of PHI to PII is 50:1 (World Privacy Forum[1]).

Attackers also have (daily) costs in order to achieve their income - "profits" are the difference between costs and potential income:

$$P = I – Kt\ ,$$

i.e.,

$$P = N_{PII}C_{PII} + N_{PHI}C_{PHI} + O – Kt\ ,$$

where:

- **P** is expected profit for attacker from target.
- **K** is the *daily cost of executing a cyber-attack* (including reconnaissance, infrastructure, outsourcing, cost of hiring insiders, paying for credentials, cost of zero-day vulnerabilities, consequences of getting caught, etc.).
- **t** is time required to breach the target.

Then, for the attacker, the aim is to maximise the profit function **P** across all targets:

$$\mathrm{Max}(\ P\ ) = \mathrm{Max}(\ N_{PII}C_{PII} + N_{PHI}C_{PHI} + O – Kt\ )\ .$$

Therefore, for the attacker to ascertain target desirability simply means sorting targets in descending order by **P**. I.e., it will be preferable to attack firms with a higher **P** (profit function) first.

Each attacker will have their own, potentially unique, list in which a firm appears at a certain percentile rank **R**. NB:

Absolute Rank = No 1 ---> Percentile Rank = 0%
Absolute Rank = No [Last] ---> Percentile Rank = 100%.

Since any one firm will be a target for multiple attackers, with various different value of **R** in each attacker's desirability list, *the sum of **R**, across all considered attackers **n**, for any one target firm, is a measure of the overall susceptibility of the target to attackers.* We therefore define PTBA, the Propensity To Be Attacked, as:

$$\text{PTBA} = (\textstyle\sum_{n} R) \,/\, n \,.$$

The higher this metric, the higher is the likelihood to be attacked (elevated risk).

3. SOME OBSERVATIONS ON KEY PARAMETERS

C (expected price per record):

- Given the sensitive nature and value of healthcare information, it is no surprise that $\mathbf{C_{PHI}} > \mathbf{C_{PII}}$[1].

O (other gains expected by attacker from target):

- Is highly correlated to the target's industry. E.g., investment banking, hedge funds[2], law firms, accounting firms, etc., all have (motivating) gains, other than data exfiltration, for an attacker. E.g., ready funds to transfer, etc.

- May be high for smaller vendors working for larger corporations: attackers can leverage such a relationship to attack the larger organisation. The Target breach was via a HVAC vendor[3].

- For hacktivists, terrorists and nation states, **O** is non-monetary. As with terrorism, their aim is often to maximise propaganda-of-the-deed than monetary profit (**P**).

- The ransom demanded from ransomware victims is an example of **O**. Generally, the ransom is designed in a way that the victim is better off paying quickly without waiting long, thus ensuring that the cost of suffering (cost to recreate data + cost of unavailability of systems) is below the ransom amount. Globally, about 40% of victims pay[4].

K (daily cost of executing a cyber-attack):

- Depends on the type of attacker. A sophisticated and well-resourced attacker capable of absorbing larger expense would generally have a better chance against larger targets. For less sophisticated adversaries, a different victim set or different attack type (with less technical complexity e.g. Ransomware, DDoS) may maximize profits[5].

- Increases significantly[6] if there are legal or financial consequences the attacker faces for its action and can be a powerful deterrent to attack.
- To minimize **K** an attacker may try to re-use the same attack components on similar firms e.g., industry peers, or companies using similar technology. For example, Target and Home Depot hacks included variants of BlackPOS malware, the Sony hack used Destroyer which had code level similarities with Shamoon, used to attack Saudi Aramco[7].
- Attackers are opportunistic. They go after easiest targets first, not wasting time where quick results are not yielded. Attackers tend to quit when their target firm exhibits strong security[12].
- The time to deter the majority of attacks is less than two days. The longer an organisation can keep the attacker from executing, the more likely the attacker will move to the next target (a parallel from the terrorism space is *target substitution*). Higher IT maturity may therefore deter attackers from pursuit of the target firm[12].
- For calculating **K**:
  - 69% of the attackers are motivated by money. On average, attackers receive $28,744 annually for every 704 hours spent on attacks[12]. This is dissimilar to terrorism, where ideology and propaganda-of-the-deed are key.
  - Attacker technology and availability is improving, enabling more attacks. Technically proficient attackers spend an average of $1,367 for specialized tools to execute attacks[12].

  If we ignore the (possibly eventual) cost of extradition or legal costs to the attacker, we can calculate an average daily cost, the aforementioned **K**:

$$K = (\ \$28{,}744 + \$1{,}367\ )\ /\ 704 \text{ hours} = \$42.8 \text{ per hour.}$$

t (time required to breach target):

- Depends on both the maturity of IT security employed by the victim and the sophistication of the attacker.

PTBA (Propensity To Be Attacked):

- Annual revenues are not an exact indicator for PTBA, since the target could be in business of managing third party data (e.g., payroll processors, accountants) which could be of a different value to what its own revenues might imply.

- A bank might have a very high desirability and a large payoff. What prevents it having a high PTBA, is that its stringent countermeasures attenuate its profit function, hence it is by no means guaranteed that a bank would be first in the percentile ranking of profit function.
- PTBA is dimensionless: regardless of how long the list of targets held/considered by each attacker, and regardless of how complete the attacker spectrum characterisation, it is a normalised score between zero and one.

4. CALCULATING EXAMPLE PTBAS

In principle, highly granular data on each individual attacker could be ascertained via the dark web and/or sinkholes. However, since representative (let alone exhaustive) compilation of these is not currently possible in practise, using *attacker groups*, a broader and more practical classification, covers all types of attacker.

Using VCDB data, we can calculate the PTBA across a range of industry classes for a spectrum of attacker types, across a two-year period (2015-2016):

Sector	PTBA			
	Crime	Hacktivist	Nation State	Malicious Insider
Accommodation	0.789	0.526	0	0.631
Administrative	0.315	0	0	0.473
Agriculture	0	0	0	0
Construction	0	0	0	0.157
Educational	0.684	0.526	0	0.842
Entertainment	0.315	0	0	0.263
Finance	0.894	0.842	0	0.894
Healthcare	1	0.842	0	1
Information	0.631	0.947	0.736	0.684
Manufacturing	0	0	0	0.526
Mining	0.315	0	0	0
Other Services	0.578	0.789	0.736	0.578
Professional	0.684	0.736	0.947	0.789
Public Sector	0.947	1	0.947	0.947
Real Estate	0	0	0	0.263
Retail	0.842	0.526	0.736	0.736
Trade	0.315	0.526	0	0.368
Transportation	0	0	0	0.421
Utilities	0.315	0	0.736	0.157

Since PTBA can be calculated for any granularity, we can also combine all attacker types to more simply look at how the threat landscape changes over time (2015-2016 to 2016-2017):

Sector	PTBA	
	2015-16	2016-17
Accommodation	0.4865	0.44425
Administrative	0.197	0.111
Agriculture	0	0
Construction	0.03925	0.097
Educational	0.513	0.5135
Entertainment	0.1445	0.49975
Finance	0.6575	0.666
Healthcare	0.7105	0.6665
Information	0.7495	0.72175
Manufacturing	0.1315	0.13875
Mining	0.07875	0.0555
Other Services	0.67025	0.6385
Professional	0.789	0.5275
Public Sector	0.96025	0.87475
Real Estate	0.06575	0
Retail	0.71	0.722
Trade	0.30225	0.208
Transportation	0.10525	0.111
Utilities	0.302	0.2775

Hence, we infer that the public sector is the most hazardous industry class, while agriculture is the least, borne out empirically, *and* according to the PTBA measure which objectively quantifies such risk.

5. CONCLUSION

We have shown that using readily available historical data, or forecasts for future events[15], that it is possible to calculate a single-number risk score: the PTBA (Propensity To Be Attacked). This metric takes into account both attacker motivations and cost, and defender cyber countermeasures. It varies correctly across time, industry, and attacker type. It is flexible and dimensionless: regardless of how long the list of targets held/considered by each attacker, and regardless of how complete the attacker spectrum characterisation, it is a normalised score between zero and one, making it ideal for underwriting both single risks and portfolios, for insurance and reinsurance.

REFERENCES

1. *2017 Cost Of Data Breach Study.* https://www.ibm.com/security/data-breach/
2. *Cyber Attackers Turn Their Focus On Hedge Funds.* http://usblogs.pwc.com/assetmanagement/cyber-attackers-turn-their-focus-on-hedge-funds/
3. *Target Hackers Broke In Via HVAC Company.* https://krebsonsecurity.com/2014/02/target-hackers-broke-in-via-hvac-company/
4. *Malwarebytes: Understanding The Depth Of The Global Ransomware Problem.* https://go.malwarebytes.com/OstermanRansomwareSurvey.html

5. *Ransomware As A Service.* https://documents.trendmicro.com/assets/resources/ransomware-as-a-service.pdf
6. *Three Chinese Hackers Fined $9 Million For Stealing Trade Secrets.* http://thehackernews.com/2017/05/chinese-hacker-trade-secrets.html
7. *Recycle, Reuse, Reharm: How Hackers Use Variants Of Known Malware To Victimize Companies And What Paypal Is Doing To Eradicate That Capability.* https://www.paypal-engineering.com/2015/11/19/recycle-reuse-reharm-how-hackers-use-variants-of-known-malware-to-victimize-companies-and-what-paypal-is-doing-to-eradicate-that-capability/
8. *Black Market Medical Record Prices Drop To Under $10, Criminals Switch To Ransomware.* http://www.csoonline.com/article/3152787/data-breach/black-market-medical-record-prices-drop-to-under-10-criminals-switch-to-ransomware.html
9. Chopitea, Thomas. 2012. *Threat Modelling Of Hacktivist Groups – Organization, Chain Of Command, And Attack Methods.* http://publications.lib.chalmers.se/records/fulltext/173222/173222.pdf
10. *Crowdstrike, Art Of Attribution* https://www.rsaconference.com/writable/presentations/file_upload/anf-t07b-the-art-of-attribution-identifying-and-pursuing-your-cyber-adversaries_final.pdf
11. *NIST Cybersecurity Framework* https://csrc.nist.gov/Projects/Program-Review-for-Information-Security-Assistance/Security-Maturity-Levels
12. *The Real Cost Of Attacks* https://media.paloaltonetworks.com/lp/ponemon/report.html
13. Holt, Thomas J and Smirnova, Olga. 2014. *Examining The Structure, Organization, And Processes Of The International Market For Stolen Data.*
14. VCDB Github, https://github.com/vz-risk/VCDB
15. Ismail, Raveem & Werner, Christoph. 2017. *Structured Expert Judgement for (Re)insurance: Forecasting Political Violence frequency.* Journal Of Terrorism & Cyber Insurance. Vol 1 No 1. https://1drv.ms/b/s!AjWDpOlDwNZGgbgsgU8u0E6BYYJ5-w

APPENDIX: WALKTHROUGH FOR REPLICATING CALCULATIONS

The PTBA formulation is agnostic to data used: here, we have used the VCDB database for its virtues of being open source and having good coverage (7,300 events at the time of writing).

The reproducible steps and assumptions are:

1. Group by attacker class (Malicious Insider, Nation State, Criminak, Hacktivist):
   - **Malicious Insider**: "actor.internal" = "TRUE". Accidental data releases do occur, but without capturing true motives, malicious intent can be assumed for conservatism.
   - **Nation State:** "actor.external.variety.Nation-state" or "actor.external.variety.State-affiliated" = "TRUE".
   - **Crime:** "actor.external.variety.Organized crime" = "TRUE".
   - **Hacktivist:** "actor.external.variety.Activist" = "TRUE".
2. Calculate the number of attacks by each actor category by year and industry class (simple pivot). We used victim.industry.name as the industry class, and timeline.incident.year as the year.
3. In the absence of further data, we assume that the number of events (for a particular attacker type and industry) is the manifestation of ranking in target desirability.
4. Calculate PTBA for a particular industry for a given year and an attacker category (our first table): $\text{PTBA}_{\text{industry}} = \text{PercentileRank}_{\text{industry}}$, where, in the absence of further data, the number of attackers is assumed the same for each attacker category.
5. For PTBAs without breaking out attacker types (our second table), we simply sum PTBAs across attacker type (each row in our first table) and divide by 4, exhibiting the additive utility of the risk score.

It should be noted that the PTBA calculation is a framework for dealing with a heterogonous spectrum of data quality. The VCDB data are extremely basic, but we have shown how they would plug into the PTBA calculation – comprehensive data on unknowable or challenging to acquire information (such as desirability, number of attackers, etc.) would go straight into the PTBA calculation and improve it, but even in its absence, a useful metric can be produced.

Enhancing the Generalized Linear Modeling Approach with Machine Learning Technique

Jie Dai, FCAS, CSPA

Abstract

With the development of the machine learning (ML) technique and broad successful application, machine learning is becoming more and more popular for data analytics in many industries. Insurance is no exception, and machine learning techniques are used to build predictive models in Claims (Fraud, subrogation models), Marketing (Segmentation, cross sell model, recommendation models), and Underwriting. However, for pricing models, Generalized Linear Models (GLM) still dominates given its easy interpretation and well-established frame work. Using a machine learning method to enhance the GLMs model is a challenge to the insurance industry especially for actuarial modeling. This paper will discuss some potential ways to enhance the GLMs model with tree based machine learning techniques and give a case study on territorial analysis, which would show significant improvement on the predictive nature of the GLM model.

Keywords. Machine learning; territorial analysis; generalized linear modeling.

1. INTRODUCTION

Machine Learning is a sub-set of artificial intelligence where computer algorithms are used to autonomously learn from data. Machine learning (ML) is getting more and more attention and is becoming increasingly popular in many other industries. Within the insurance industry, there is more application of ML regarding the claims and underwriting disciplines. There is little in actuarial literature on ML, and none is in pricing modeling.

In the early 1970s, Nelder and Wedderburn coined the term generalized linear models (GLM) for an entire class of statistical learning methods that include both linear and logistic regression as special cases. In the last two decades, GLMs have been widely in use in P&C insurance to classify risks and determine rate structures. However, standard GLMs do have several shortcomings, most notably [1]:

- Predictions must be based on a linear function of the predictors;
- GLMs exhibit instability in the face of thin data or highly correlated predictors;
- Full credibility is given to the data for each coefficient, with no regard to the thinness on which it is based;
- GLMs assume the random component of the outcome is uncorrelated among risks;
- The exponential family parameter $\emptyset$ must be held constant across risks;
- GLMs only can identify simple and global interactions, which are the interactions between

all levels of two predictors. For identifying complex interactions with GLMs, the manual process would be non-trivial.

Also, another challenge of using GLMs includes the selection of predictors from large volume of variables candidates.

In mid 1980s Breiman, Friedman, Olshen and Stone introduced classification and regression trees, which for the first time made fitting non-linear relationships computationally feasible. Since then there are more algorithms (like neural nets, random forests or gradient boosting) that have been developed and widely used in other industries or disciplines [2]. Those methods don't have the shortcomings noted above, and therefore able to produce strong models that have the potential to yield more accurate predictions. However, using those methods directly would entail a huge loss of interpretability, which is critical for many actuarial applications.

This paper will present the ways to enhance the GLMs with ML technique in variable selection and feature engineering. In addition, we will look at an application in sewer backup modeling that shows significant improvement of the model results with the new features created through ML. However, for reasons of confidentiality, we are not able to share detailed data and quantitative results in this paper.

1.1 Research Context

With more and more data being available for pricing models, the challenge arises to reduce the number of predictors to improve the prediction accuracy and interpretability. Stepwise selection (forward, backward and/or hybrid) are widely used in GLM modeling practice. Recently, shrinkage methods like Lasso (least absolute shrinkage and selection operator) have become more popular because it can be a more efficient method that produces more interpretable models that involve only a subset of the predictors. The third method to reduce variables or dimensions is to create predictors from the original raw predictors. Principal Components Analysis (PCA) is the most popular approach in deriving a low-dimensional set of features from a large set of variables. Insurance score is a major rating variable introduced to personal lines insurance [3] in the late 1980's and 1990's. This variable is derived from dozens of selected/created credit variables (from initially thousands raw variables) to predict insurance loss risk by using linear regression and/or ML. Another popular rating variable in auto insurance which is created from dozens of raw vehicle characteristic variables is auto symbol. Both variables are easier to interpret and reduce the dimension significantly compared to using the raw underlying variables.

Interaction identification is a challenge in GLMs modeling in practice, especially for the interaction among more than 3 variables. Some ML techniques naturally will include all the possible interactions between variables. Creating new features based on the ML techniques to replace the underlying raw variables would not only reduce the number of variables but also significantly improve the predictive power of the GLMs model.

To do the territorial analysis for a sewer backup modeling, 14 geographic variables are studied which are not predictive in the model. A score variable was created from these 14 variables, and a territorial definition was created from census block group with the help of the 14 geographic variables. The score variable can be used in underwriting and pricing. Both new features would improve the predictiveness of the GLM model significantly.

1.2 Objective

Our objective is to use the feature created with ML from some underlying variables to improve the predictive power of the GLMs. Those new features should be like vehicle symbol or credit score which can be interpreted to a certain degree.

1.3 Outline

The reminder of the paper proceeds as follows. In Section (2.1), we discuss the sewer backup data and modeling. In section (2.2), we discuss the territorial analysis, and especially the challenge for sewer backup loss data. In section (2.3), we discuss the tree based supervised learning methods in ML. In section (2.4) we introduce the double lift curve for the model comparison. In section (3.1) we present the result that shows even if the raw variables are not good predictors, the score produced from them through ML can be very predictive. Finally, in section (3.2) we present the model comparison with and without the boundary, which shows the significant improvement with the boundary variable. The boundary variable is created by grouping census block group.

2. BACKGROUND AND METHODS

2.1 Sewer Backup Modeling

The sewer backup loss modeling dataset included observations with sewer backup coverage endorsement. Since this loss is highly correlated with location, the territorial analysis should be important. To do the territorial analysis and create the boundary, we tested 14 geographic variables from US census data. The 14 geographic variables include Water Surface Elevation, Average Travel Time, Average Household Size, Average Number of Vehicle, Population Growth in 5 years, Average Age etc. Unfortunately, none of them showed predictive power. It also is difficult to create a territorial boundary with a spatial smoothing method. For this study, we only present the result for frequency models.

2.2 Territorial Ratemaking and Boundary

For territorial ratemaking, the first phase is to establish territorial boundaries [4]. Census block group (CBG) is selected as the basic geographic unit due to its small size and relative stasis over time. The current approach to create the boundaries include the following steps [4]:

- Create geographic estimator on CBG with geographic indictors by building a GLM model using a variety of non-geographic and geographic explanatory variables;
- Applies spatial smoothing techniques to the geographic residuals to see if there are any patterns in the residuals and those residuals can be used to adjust the geographic estimators to improve overall predictive power of the model.
- Once the geographic estimators are calculated for each CBG, the CBG can be grouped into territories.

Our proposed approach is to create the CBG estimator by building a GBM (gradient boosting machine) model on the residual of the GLM model by using the 14 geographic variables. The GLM model is created using non-geographic and geographic explanatory variables. With the help of smooth weight of evidence (SWOE) [5] we transferred the categorical variables (CBG) into an interval variable, and then created a boundary based on the decision tree, we grouped the census block group into 19 levels.

2.3 Tree-Based ML Techniques

Tree based methods partition the feature space into a set of rectangles, and fit a simple model (like a constant) in each one [6]. Assume our data consists of p inputs and a response, for each of N observations: $(x_{i,}\ y_{i,})$ for i=1,2,…N, with $x_i\ = (x_{i1,}x_{i2,\ldots,}x_{ip,})$. For regression tree, if we have a partition into M regions R_1, $R_2\ldots\ R_M$, and we model the response as a constant C_m in each region:

$$f(x) = \sum_{m=1}^{M} c_m\, I(x\epsilon R_m\ ) \qquad (2.1)$$

It is easy to see that the best $\hat{c}_m$ is just the average of $\mathrm{y_i}$ in region R_m :

$$\hat{c}_m = ave(y_{i,}|x\epsilon R_m). \qquad (2.2)$$

The big advantage of a tree based ML technique is that it is easy to interpret, and easy to implement. It is still a great tool for identifying interaction or as a supplement analytic tool for other more advanced techniques. One major problem with trees is their high variance [6]. A small change in the data can result in a very different series of splits, making model chosen somewhat precarious. To reduce this variance, several tree based algorithms have been developed, which are more predictive and would reduce the possibility of over fitting the model. Among them, the two most common of these techniques used are boosting and bagging.

A Gradient Boosting Machine (GBM) is a generalization of tree boosting that attempts to mitigate some problems with other boosting methods like speed, robustness and interpretability [6]. The generic algorithm for the GBM is listed here [6]:

```
$Initialize\ f_{0(x)} = argmin_{\gamma} \sum_{i=1}^{N} L(y_i, \gamma)$
For m=1 to M:
    For i=1,2,…N compute $r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}$
    Fit a regression tree to the targets $r_{im}$
    For i=1,2,… $J_m$ compute $\gamma_{jm} = argmin_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$
```

$$f_m(x) = f_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm}) \tag{2.3}$$

Where $L(y_i, \gamma)$ is the loss function, and the parameter ν can be regarded as controlling the learning rate of the boosting procedure. Both ν and M control prediction risk on the training dataset. Smaller values of ν lead to larger values of M for the same training dataset, so that there is a tradeoff between them. When M is large, the computation becomes expensive and would take a long time to run. To our experience, ν may vary from 0.01 to 0.15 and M can be from 50 to hundreds depending on the dataset. We run the model with SAS enterprise miner, other tools or package (R or Python) of gradient boosting may choose different ν and M to get the best result.

Random forest is a substantial modification of bagging that builds a large collection of de-correlated trees, and then averages them. On many problems, the performance of random forests is very like boosting, and they are simpler to train and tune, and random forest is easier to parallelize and robust to overfitting. That's why random forest is also popular in ML application. However, in the author's experience, we do see GBM outperform random forest in many insurance applications if it is well tuned.

The advantages of tree based models over GLM include but are not limited to:

- No assumption of model structure which would be learnt from data;
- Easy implementation of complex and/or multiple way interactions;
- Easy to deal with missing values;
- Built-in feature selection;

2.4 Double Lift Curve

For modeling comparison, a double lift curve is a simple method to directly compare the predictive accuracy of two models. Here we use EMBLEM's model comparison function to compare two model's performances. The X axis is the bucketed ratio of indications of the two models, and the graphs will show the two models' average indications in those buckets and the average of actual observations in those buckets. The "winning" model would be the one that matches better the observed frequency in each bucket. In all the following models, we split the dataset into 80% and 20% randomly as training and validation dataset, and the double lift curves are created on the validation dataset using EMBLEM.

3. RESULTS AND DISCUSSION

3.1 Geographic Variable Score

To show the idea that the complicated interactions are important and are missed sometimes in the GLM modeling, we built two Frequency models: model1 is the model with all the current rating variables plus 14 geographic variables; model2 is the model with all the current rating variables plus a geographic variable score (geoonly14), which is created based on the 14 geographic variables with GBM.

Fig 3.1 shows the double lift curves for model 1 and 2. Based on these results, we see that model 2 is significantly better in predictive accuracy. This result shows a case where even when the individual variables are NOT predictive; the combination of the variables can be very predictive because of the complicated interactions between those underlying variables. Looking for interactions among 14 variables with many levels would be a non-trivial work and especially difficult because we generally have no prior knowledge regarding the potential interactions between geographic variables. And we are not able to identify/include interactions among 3 or more geographic variables in GLM with EMBLEM.

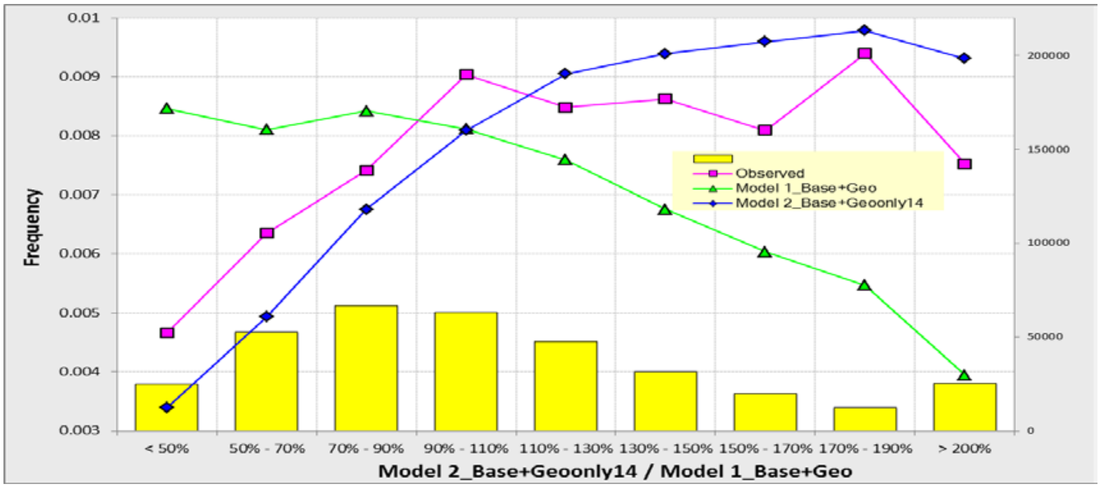


Fig 3.1 Double Lift Curve for the Model Comparison for Geo Variables and Score

3.2 Sewer Backup Territorial Boundary

For territorial analysis, the current method is to use geographic variables to create the indication for the census block group (CBG) with GLM modeling, and then use the classifier of EMBLEM to do the spatial smoothing and correction (if there is pattern in the residuals). However, it is very difficult to find the pattern in the residuals, and thus the correction is also subjective in practice. In theory, we can use CBG as the variable to create the indication for territorial rating. The hurdle would be how to group the more than ten thousand levels of CBG. We use SWOE to recode the CBG and with the help of a decision tree model on the GBM model output, we can create the CBG group which could be used in the territorial rating directly. We produced the 19 CBG groups and incorporated it into the base model for our sewer backup classification GLM model. Fig 3.2 shows the comparison of the two models: Model 1: Base model (current rating plan) plus the geographic variables; Model 2: Base model plus the 19 CBG grouping variable (Territorial Boundary). Model 2 shows significant improvement over Model 1 in predictive accuracy.

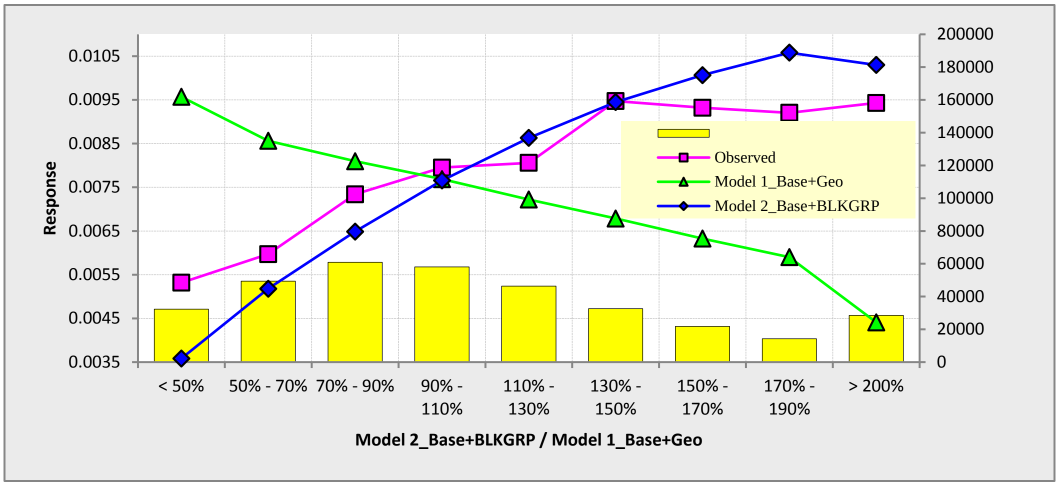


Fig 3.2 Double Lift Curve for the Model Comparison for Geo Variables and Territorial Boundary

4. CONCLUSIONS

With the development of the advanced modeling techniques, there are more and more data and variables available for pricing. It is a challenge to select variables and/or extract information from those raw variables to build a model which is more accurate in predictive power and still interpretable. To keep the GLM framework intact, the methods presented in this paper show the potential ways to incorporate advanced analytical techniques, especially machine learning, into the variable selection and dimension reduction procedure, which may significantly increase the predictive power of the model. This method can be applied to develop vehicle symbol, territorial boundary and other risk score variables.

5. REFERENCES

[1] Mark Goldburd, Annand Khare and Dan Tevet, "Generalized Linear Models for Insurance Rating", CAS Monograph Series Number 5.

[2] Gareth. James, Daniela Witten, Trevor J. Hastie and Robert John Tibshirani., An Introduction to Statistical Learning, Springer Science & Business Media, 2013.

[3] Conning Report, "Insurance Scoring in Personal Automobile Insurance – Breaking the Silence", Conning Report, Conning, (2001).

[4] Geoff Werner, Claudine Modlin, Basic Ratemaking, Fifth Edition, May 2016.

[5] Satish Garla, Goutam Chakraborty, Andrew Cathie, "Extension Nodes to the Rescue of the Curse of Dimensionality via Weight of Evidence (WOE) Recoding", SAS Global Forum 2013.

[6] Trevor J. Hastie, Robert John Tibshirani, and Jerome H. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer series in statistics. Springer, New York, 2 editions, 2009.

Abbreviations and notations

CBG, census block group
GLM, generalized linear models
GBM, gradient boosting machine
ML, machine learning
SWOE, smooth weight of evidence

Biography of the Author

Jie DAI is Associate actuary at Sentry Insurance Company in Middleton, WI. He is responsible for nonstandard auto modeling. He has a degree in aerodynamics from the Northwestern Polytechnic University in China. He is a Fellow of the CAS.

Ratemaking for a New Territory: Enhancing GLM Pricing Model with a Bayesian Analysis

Jing Zhang and Tatjana Miljkovic

Abstract

Motivation. This paper offers a Bayesian approach in ratemaking for a new territory where a company considers starting a new business, or for a relatively new territory where the company has very limited claims experience.

Method. A Bayesian Poisson regression model with power priors and weakly informative priors for the model parameters is proposed for modeling claims frequency. Bayesian analysis of claim severity considers a gamma regression and non-informative uniform priors for the regression coefficients.

Results. After incorporating the external information from a similar book of business in a similar territory, Bayesian analysis with power priors improved the prediction reporting a small Means Squared Prediction Error (MSPE).

Conclusions. Bayesian analysis with power priors can be used effectively in auto insurance ratemaking for pricing of a new business in a new territory, or improving pricing of a growing business in a new territory.

Availability. The original SAS code will be available for distribution pending the acceptance of this paper.

Keywords. Bayesian analysis, GLM, new territory, power priors, predictive modeling, ratemaking.

1. INTRODUCTION

1.1 Research Context

Ratemaking for a new line of business or a new territory is subject to a judgement under uncertainty. Actuaries in these situations often rely on the availability of external industry data or experience from a similar line of business, as both of these serve as heuristic benchmarks, but sometimes they lead to severe and systematic errors. If the volume of claims experience is subject to significant changes (e.g., due to catastrophic events or regulatory conditions), these estimates will be severely biased. The company may gauge some prior information about the prospective new business in a new territory by pooling this information from the existing business, assuming the new underwriting practices in a new territory will remain more or less similar to the existing underwriting practices to the territory from which this information is drawn. A new territory may also share some common demographic, geographic, or climate characteristics with one of the existing territories so that the information contained in the existing business can be utilized in the rating process of the new territory.

According to Chen and Ibrahim (2006, pp. 551), "Power priors have emerged recently as useful informative priors for the incorporation of historical data in a Bayesian analysis" and are well-received in statistical practice. These priors can be efficiently incorporated in Bayesian analysis with generalized linear models (GLM) and help incorporate useful prior information from existing territories in the context of analyzing limited information from the new territory of interest.

Most of the insurance companies are moving away from the one-way premium calculation approach by employing GLMs with the original statistical framework discussed in the book by McCullagh and Nelder (1989). The GLM models are praised for two major advantages over ordinary linear models. First, the GLMs work with a number of discrete distributions and continuous distributions, which make them more flexible compared to the ordinary linear model that is constrained by the normal distribution only. Second, the GLMs allow for some transformation of the mean as a linear function of the covariates, with additive and multiplicative models as special cases. For more extensive theory behind non-life insurance pricing using GLMs, we refer the reader to books by Kaas et al. (2008) and Ohlsson and Johansson (2010).

A frequentist approach to predictive modeling based on GLM models has the capability to predict outcomes that best represent the company's data with insufficient regard for prior probability. The probability distributions of the parameters considered in this type of modeling rely on the sampling distributions that are based on all possible random samples of experiences that could have occurred, but they are not conditional on the actual sample that did occur. A Bayesian point of view considers inferences based on the probabilities calculated from the posterior distribution, making them conditional on the sample that actually did occur. The role of prior distributions in the Bayesian analysis is to capture "pre-data" information about the parameters, then use the prior experience that was collected to update the "pre-data" information about the parameters to "posterior" information about the parameters. Thus, the Bayesian approach considers parameters as random variables.

Recently, Bayesian methods have been actively discussed in the area of predictive modeling and ratemaking. Boucher et al. (2008) used Bayesian and frequentist models based on generalization of Poisson and negative binomial distributions to account for correlation between contracts of the same insureds. The authors showed that the models based on time dependence covariates (e.g., past experience) cannot be used in modeling of reported claims. They recommended use of random effects models in computing the next year's premium as these models show improved fit compared to other models. The same authors, Boucher et al. (2009), extended their study by considering the relationship between number of accidents and number of claims using the generalization of the zero-inflated Poisson (ZIP) distribution. The authors proposed an approximation of the number of accidents distribution that can be used to provide insightful information about the behavior of insureds using panel count data. A Bayesian analysis was used in computation of the predictive distribution for the random effects.

Bermúdez and Karlis (2011) examined Bayesian multivariate Poisson models and their zero-inflated extensions for improving current ratemaking procedures. Brown and Buckley (2015) used a Bayesian approach to determine the number of groups in an insurance portfolio. The claim count is assumed to follow a Poisson distribution.

We consider the following scenario for pricing new business in a new territory, where there is no prior claims experience. First, we can identify a similar territory from our existing book of business for which the claim experience is established. These two territories may be neighbors that share similar climate, geography, and demographic characteristics. For pricing the new business during the first year with no data, we can borrow the information from the existing territory and set the new rates. After the first year, for pricing the business during the second year, we can borrow the experience from the similar existing territory in the analysis of the limited claim experience in the new territory. Then, we can run the proposed Bayesian model with power priors. We repeat this process for several years until we accumulate the claims experience in the new territory to be able to use the standard pricing method. The flow chart of this process is outlined in Figure 1. Our proposed Bayesian model with power priors would provide a new way of pricing the business for a new territory (framed part of Figure 1) and serves as the main contribution of this paper. The example that we provided in the subsequent sections would help the practitioners in implementation of this proposed method.

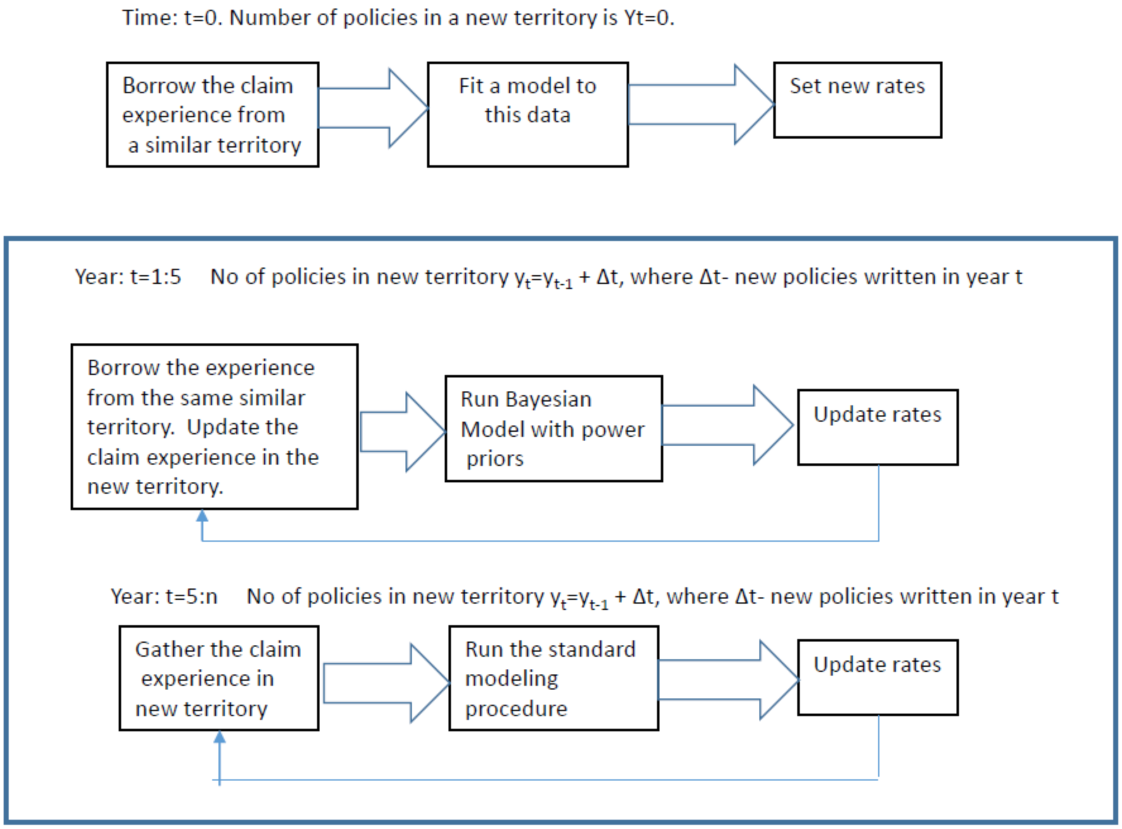

Time: t=0. Number of policies in a new territory is Yt=0.
Borrow the claim experience from a similar territory
Fit a model to this data
Set new rates
Year: t=1:5 No of policies in new territory $y_t=y_{t-1}$ + Δt, where Δt- new policies written in year t
Borrow the experience from the same similar territory. Update the claim experience in the new territory.
Run Bayesian Model with power priors
Update rates
Year: t=5:n No of policies in new territory $y_t=y_{t-1}$ + Δt, where Δt- new policies written in year t
Gather the claim experience in new territory
Run the standard modeling procedure
Update rates

Figure 1: Flow chart of the proposed Bayesian method for pricing in a new territory.

1.2 Objective

The objective of this paper is to introduce a Bayesian approach with power priors and weakly informative priors to be used in developing frequency distribution of claims for a new territory where the company has very limited experience. The historical information can be borrowed from an adjacent territory based on geographic and demographic profiles, for purpose of the Bayesian analysis. A Bayesian analysis with non-informative priors for modeling severity of claims is also illustrated in a new territory.

To our knowledge, the Bayesian GLM claim models with a Poisson distribution have not previously been considered, either with power priors or weakly informative priors. We would like to close this gap in the actuarial literature by proposing the Bayesian frequency models that use power priors and weak informative priors of the regression coefficients. This approach is especially appealing for determining the premium rates in a new territory that lacks claims experience.

1.3 Outline

The remainder of the paper proceeds as follows. Section 2 presents the Bayesian methodology for modeling frequency and severity of claims. Section 3 describes the analysis of real data and the results. Section 4 provides the summary of the model validation. The conclusion is provided in Section 5.

2. BACKGROUND AND METHODS

In this section, we explore the models for claims frequency and claims severity. For each model, we show frequentist and Bayesian approaches from a theoretical perspective.

2.1 Models for Claims Frequency

It is popular to assume that the number of claims follows a Poisson distribution and, hence, a generalized linear regression can be fitted to analyze the relationship between the number of claims and the relevant predictors.

$$Y_i|\theta_i \sim independent\ Poisson(E_i\,\theta_i\,) \tag{2.1}$$

$$\log(\theta_i\,) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}, \tag{2.2}$$

where the vertical bar "|" describes the distribution of the quantity to the left of the "|", given information to the right. Here Y_i denotes the number of claims filed by the i*th* policy holder. Here, the vector of predictors is defined as $x_i = (x_{1i}, \ldots, x_{pi})'$. The Poisson mean, $E_i\,\theta_i$, is determined by the known length of insured time (E_i , also known as the offset) and rate of claims (θ_i). Here the rate of claims is modeled as a function of the relevant predictors (θ_i), including demographic information of drivers, descriptive information of cars and residential areas. The regression coefficients, β_0, …., β_p, relate the rate of claims with these predictors.

In the frequentist approach, point estimation of model parameters can be implemented via Maximum Likelihood Estimation (MLE) or Restricted Maximum Likelihood (REML) approaches, and inferences can be made based on large sample distributions of the point estimators. Other distributional assumptions of the claims frequency can be used, such as zero-inflated Poisson or negative binomial. Initially, we considered regression models assuming these distributions as well; however, the fit of Poisson regression turns out to be the best for the data analyzed. Since the main purpose of the present study is to illustrate the incorporation of prior information from an external existing territory via Bayesian modeling of claims frequency when sample size is limited, we decided to stay with the Poisson distributional assumption.

The Bayesian analysis treats the parameters as unknown random variables. To implement the analysis, we need to propose a "prior distribution" for the model parameters. Combining the data likelihood and prior distribution of parameters using Bayes theorem, we are able to update the knowledge about the distribution of model parameters, and the updated knowledge is called "posterior distribution." The posterior distributions are then used for Bayesian inference. Here we begin the Bayesian analysis assuming independent normal prior distributions for the regression coefficients, i.e.,

$$\pi\left(\beta_j|\beta_j^0,\sigma_j^2\right) = N(\beta_j^0,\sigma_j^2),\ j = 0,1,\dots,p. \tag{2.3}$$

Higher level priors are then assumed for prior mean β_j^0 and prior variance σ_j^2 as follows:

$$\pi(\beta_0^0) = N(0,10) \tag{2.4}$$

$$\pi\left(\beta_j^0\right) = N(0,4),\ j = 1,\dots,p. \tag{2.5}$$

$$\pi(\sigma_0) = Uniform(0,5), \pi\left(\sigma_j\right) = Uniform(0,1), j = 1,\dots,p. \tag{2.6}$$

The hyper-parameters are chosen to incorporate weak informative prior distributions on the parameters. Besides the weakly informative priors, we also illustrate the incorporation of prior information from external data of similar region via power priors. The power priors have been proposed in Ibrahim and Chen (2000), with applications in hierarchical modeling discussed in Chen and Ibrahim (2006) and well-received in statistical practice.

The power prior of model parameters is constructed by raising the likelihood based on the external data to a suitable power and then multiplied by an initial prior (usually non-informative or weakly informative); therefore, power prior uses the external data with a discount relative to the data of interest, which allows a discrepancy between insurance policy holders in this similar region and the current region of interest. The power prior is a useful tool to borrow strength from external data in Bayesian analysis. In the present study, we considered a second Bayesian analysis that incorporates the external data using power prior with power of 0.5, which implies a 50% discount of external information in the log-likelihood function of the joint posterior density function of model parameters; the priors used in the first Bayesian analysis (i.e. Equations (2.4)-(2.6)) are used as initial priors in this analysis.

2.2 Models for Severity

Besides the modeling of frequency of claims, it is also of interest to study whether and how the amount of each claim (severity) is related to the relevant factors (e.g., driver's age, gas type, etc.). Claim amounts are continuous measurements and can be analyzed with ordinary linear regression or generalized linear regression (e.g., log-normal regression or gamma regression). Note that the distributional assumptions that allow heavier right tails are usually a better fit to the loss data due to right-skewness of such data. When claim amounts are assumed to follow gamma distributions,

$$Z_i|\mu_i, \nu_i \sim Gamma(\mu_i, \nu_i) \tag{2.7}$$

Or equivalently,

$$f(Z_i|\mu_i, \nu_i) = \frac{1}{\Gamma(\nu_i)}\left(\frac{\nu_i}{\mu_i}\right)^{\nu_i}(z_i)^{\nu_i-1}\exp(-\frac{\nu_i z_i}{\mu_i}) \tag{2.8}$$

where ν_i is the shape parameter of the gamma distribution, and μ_i is the mean of the gamma variable and relates the covariates with the severity response. Using a log-link function, we have.

$$\log(\mu_i) = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \cdots + \gamma_p x_{pi} \tag{2.9}$$

The regression coefficients, γ_0, …., γ_p, relate the severity of claims with the set of predictors defined as $x_i = \left(x_{1i}, \dots, x_{pi}\right)$. Note that we assumed the same set of predictors are considered in the analysis of frequency and severity of claims in Equations (2) and (9), which can be modified in practice according to availability of data and prior beliefs. The two sets of covariates used in these two models are not necessarily the same.

Frequentist approaches can be used to fit these generalized linear models described above to the severity data of the new territory, and likelihood-based inference would help us determine the relationship between severity and covariates. When expert knowledge or existing analysis results concerning this relationship from a similar territory are available, the Bayesian approach would help us incorporate the information through prior elicitation. However, we believe that one should be cautious of using power priors in the analysis of severity since the potential outliers or heavy right tail in the severity observations of the "external" data might introduce misleading information in the analysis and bias the conclusions. In the present study, we used non-informative uniform priors for the regression coefficients:

$$\pi(\gamma_i) \propto 1, i = 0, 1, \dots, p. \tag{2.10}$$

The prior distribution of shape parameter is specified through the following parameterization. Let $\kappa_i = \frac{\nu_i}{\mu_i}$ be the rate parameter, then we assume an inverse-gamma prior distribution for the rate parameter as follows,

$$k_i \sim InverseGamma(0.001, 0.001) \tag{2.11}$$

The specified prior distributions would then provide vague prior input for the analysis.

3. RESULTS AND DISCUSSION

In this section, we illustrate the proposed methodology using the data from a French insurance company, related to 677,991 motor third-party liability policies. The data set includes exposure information as well as the loss information and can be found as part of the "CASdatasets" library in the R software (CASdatasets). The discussion about the datasets used in the book by Charpentier (2014) and the book itself, can be found in the book review by Miljkovic (2017). Charpentier (2014) discussed the modeling of claims frequency and severity of this data based on a frequentist approach, using various GLM models. The rating factors include: region (R11, R23, R24, R25, R31, R53, R54, R72, R74), car age (0-100), density (2-27000), engine power (12 levels), brand (7 types), driver age (17-99), gas type (2 levels), and exposures in years (0.003-1.990).

In order to illustrate our methodology, we randomly sampled 1000 policies from the region R24 with density between 200-4500. This is the largest region in France that accounts for 39% of the business written. Miljkovic and Fernández (2018) used the policies from the same region (R24) to illustrate how the unobserved heterogeneity can be modeled in an insurance portfolio using two different mixture-based clustering approaches. The histogram of the number of claims in this region as well as the severity of the claims are shown in Figure 2. The frequency of claims in this regions is: 96.3% of zero claims, 3.5% of single claims, and 0.2% of two claims. Figure 2 also shows the density of the severity of claims in R24. Minimum claim amount in this region is 2 while maximum amount is 2,036, 833 Euros. Skewness coefficient of the claim severity data is 75.12.

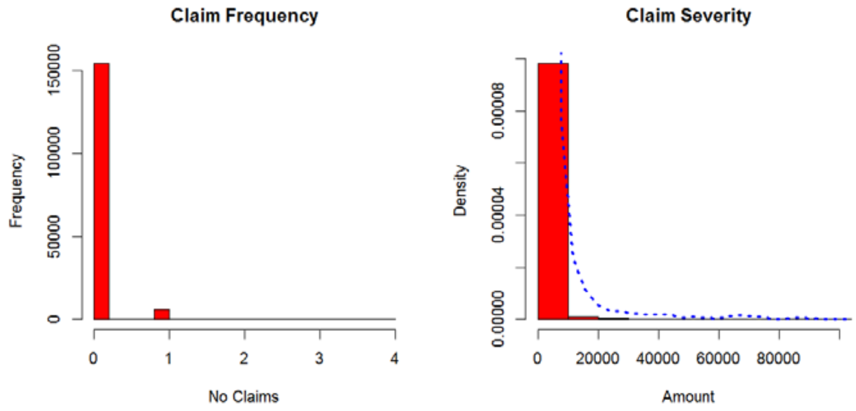


Figure 2: Frequency of claims (left) and severity of claims (right) in R24.

Our random stratified sample of 1000 policies maintains the same characteristics of R24 based on the number of policies, gas type, density, and driver's age. Gas type has two levels, diesel and regular, with regular treated as a base level. Driver age is grouped at five levels: (17-20] (base level), (20-26], (26-42], (42-74], and 74+. Density is treated as a continuous predictor. The same variables have been used by Joan-Philippe and Arthur Charpentier (Charpentier, 2014) when modeling the same data set using Poisson and Negative Binomial regression. In the analysis of frequency or severity, we standardized the density variable since it is fairly big in numerical values and results in a numerical problem in model fitting if we use the raw measurement. The new standardized density measurements are the raw density measurements subtracted by the mean density and then divided by the standard deviation of density measurements. We used "PROC STANDARD" in SAS to assist the standardization of this variable.

Both the frequentist and Bayesian analysis are implemented with SAS with the SAS code included in Appendix B. The frequentist Poisson regression model fit was obtained via "PROC GENMOD," while the Bayesian model fit of claims frequencies was obtained via "PROC MCMC." In the analysis with the Bayesian Poisson regression model assuming weakly informative priors or power priors, 20,000 samples of parameters are simulated from the posterior distributions using Markov Chain Monte Carlo (MCMC) algorithm, which are obtained from 650,000 MCMC iterations with the first 150,000 cycles as burn-in iterations and a thinning rate of 10 (i.e., every 10th draw from the MCMC simulation is used to compute credible sets and medians of the posterior distribution).

The frequentist and Bayesian gamma regression fit was obtained via "PROC GENMOD" while the Bayesian analysis utilized the "Bayes" statement provided in "PROC GENMOD." In the Bayesian analysis of claim severity, 10,000 posterior samples are obtained from 12,000 MCMC iterations with the first 2,000 cycles as burn-in iterations and a thinning rate of 1 (i.e., no thinning was used here). Convergence of the posterior simulation was evaluated using history plots and autocorrelation (ACF) plots of the posterior samples. Figures of the posterior sample of regression coefficients are shown in the Appendix A (Figures 4-6). All of the history plots show that the posterior simulation achieved convergence, while the ACF plots show that the (thinned) posterior samples do not have strong autocorrelation.

Table-1 in the Appendix A shows the comparison of the results of the Poisson GLM regression model that has been run using a frequentist approach and a Bayesian hierarchical modeling approach with weakly informative priors and power priors. For each of these three methods we show the coefficient estimates with their standard errors and the 95% confidence intervals. In the Poisson model, all of the coefficient for driver age are negative relative to the base group (17, 20] with the largest coefficient reported for age group (26, 42]. Thus, this age group reports on average the lowest frequency of claims relative to age group (17, 20]. These results are in line with other studies showing that young drivers (17, 20] are most likely to get into car accidents. The coefficient for regular gas type is negative relative to diesel gas type. The coefficient for population density in R24 is positive, indicating that an increase in population density results in additional claims reported.

From Gamma regression model, we observe that the coefficients for age group (20, 26] and (26, 42] are negative relative to the age group (17, 20] indicating that severity of claims for these groups is lower compared to group (17, 20]. The coefficients for age groups (42, 74] and (74+) are positive relative to age group (17, 20]. Also coefficient for density variable is positive indicating that the severity of the claims will increase on average as the population density increases.

Since the power prior is expressed as a product of the weighted likelihood of parameters, conditional on the historical information and a prior distribution of the parameters before the data are observed, a scale or discounting parameter from 0 to 1 is used to control the weight assigned to historical data. This parameter is usually controlled by user. Our Bayesian model with power priors assumes that 50% of external information is incorporated in the posterior distribution in the form of a prior input consisting of 50% of the log-likelihood of these external territory observations; thus, the scale parameter is 0.5.

We observe that standard errors of the posterior estimates are smaller compared to those generated with the ordinary GLM. As a result, the 95% confidence intervals are narrower than those produced with ordinary GLM or the Bayesian GLM with non-informative priors. Poisson regression results arrive at the same conclusion in terms of the risk associated with all age groups compared to age group 17-20. However, the smaller confidence intervals indicate the improvement in the estimation of the likelihood by using past information. Power priors allow for a different percent of external information to be used, which allows an actuary to judgmentally incorporate this aspect of modeling into the analysis. Another sample of 1000 losses was selected out of 16,181 policies that reported positive claim amounts. Table 2 shows the comparison of the results of Bayesian gamma regression with non-informative priors to those produced using the frequentist approach. We can also observe that standard errors and the 95% confidence intervals related to the regression coefficients are smaller compared to those produced using the frequentist approach.

4. MODEL VALIDATION

Model validation is an important part of model building. When two competing models are evaluated, common techniques such as Receiver Operating Characteristic (ROC) Curves or Double Lift Charts can be used. These techniques are appropriate, e.g logistic regression models, and require that a database of historical observations is augmented with the predictions from each of the competing models (Goldburd at al., 2016). Considering the nature of our application, the historical database is not available in a new territory where the company starts writing new business for the first time, or to an existing territory where the new business was recently introduced, so the claims experience is very limited. In absence of the historical database, we borrowed the information from the "imaginary" adjacent territory that we assumed to be R24.

Our validation is based on the "splitting data" approach and it is shown in the flowchart in Figure 3. This approach assumes drawing three samples from R24:

1) Training Set - used to perform the model building,
2) Holdout Set (Test Set) - used to perform data validation, and
3) The Bayesian "External Prior" Set - used to provide prior input information.

Both the Training Set and the "External Prior" Set consist of 1000 observations, while the Test Set consists of 100 observations. The comparison was done to evaluate the impact of incorporating the information from existing external territories on the Bayesian analysis of the Training Set. Table 3 in Appendix A summarizes the results of this validation. The Bayesian analysis with weakly information priors was applied to fit the Training Set and the predicted numbers of claims for the Test Set observations were obtained based on the corresponding posterior predictive distributions. Then we also fit the Bayesian analysis with power prior information from the External Prior Set to the Training Set and obtained the predicted number of claims for the Test Set using the new posterior prediction distributions. The two sets of predicted number of claims are both compared with the original observed frequencies for the Test Set and MSPEs were computed: 0.51 for the Bayesian analysis with weakly information priors and 0.49 for the Bayesian analysis with power priors. The MSPE calculation includes three values based on frequency of claims shown in Figure 2.

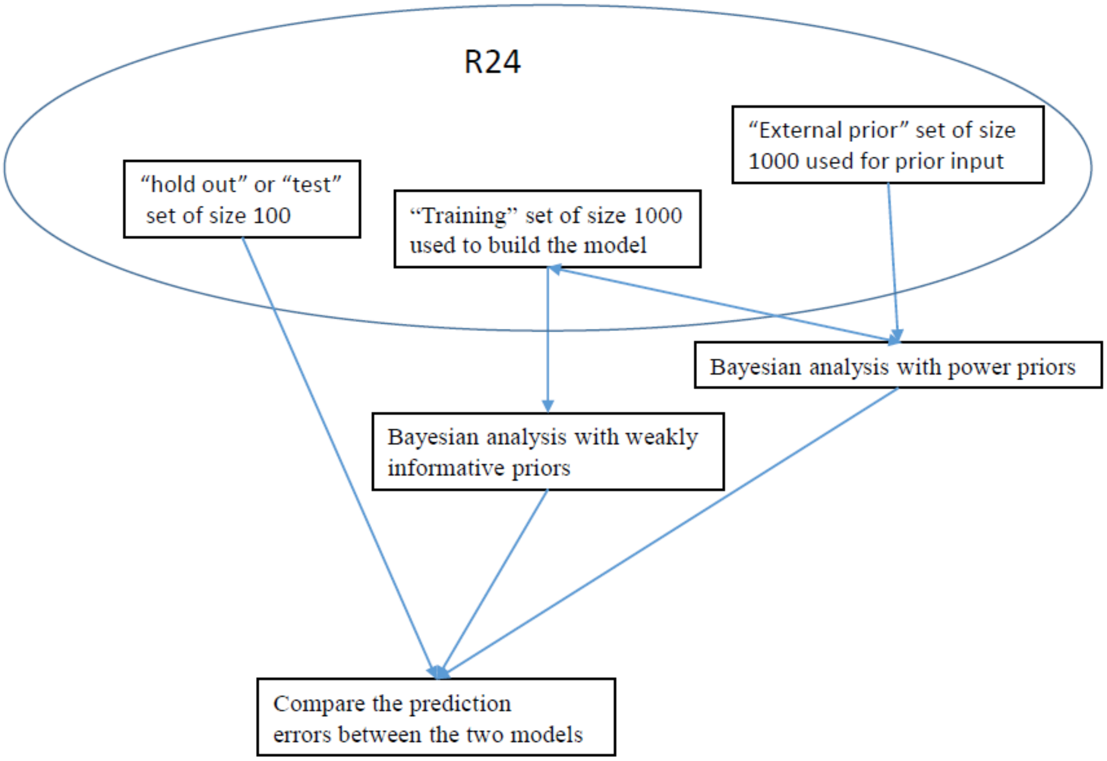

R24
"hold out" or "test" set of size 100
"Training" set of size 1000 used to build the model
"External prior" set of size 1000 used for prior input
Bayesian analysis with power priors
Bayesian analysis with weakly informative priors
Compare the prediction errors between the two models

Figure 3: Flow chart of the validation process.

We also fit frequentist Poisson regression on all three samples respectively to check the similarity of the training data, validation data, and external prior information. This validation analysis indicates better prediction performance when power priors are used. However, the strength of improvement when the power priors are used in the Bayesian, is subjected to two critical factors:

(1) Sample size of the "current" data (i.e. Training Set in this validation analysis). When sample size is fairly high relative to the complexity of the model fit, the information borrowing through power priors would play a minor role in the model prediction performance.

(2) Similarity between the External Prior Set and "current" data. When the information borrowed through power priors are "misleading", it is not favorable to incorporate those in the analysis.

5. CONCLUSION

Historical claims information is available to the actuary for the purpose of ratemaking in those territories where the company has been writing business for some period of time, even when no information is available in the case of a new territory where a company is entering business for the first time. The understanding of the "new territory" can almost always be augmented by existing information. Such borrowing of strength from historical data has long been encouraged in many scientific fields. These issues motivated us to investigate the feasibility of a Bayesian power prior approach in borrowing of strength for modeling auto rates in a new territory. The goal of such an approach is to determine a practical amount of strength to borrow from the historical claims that strikes a balance between increased cost-efficiency and long-run statistical integrity. The methods for incorporating historical data should be robust to prior knowledge and consistent with the accumulating historical information. We aim to utilize historical information given strong evidence that this information would apply well in a territory that shares some common characteristics. A more attractive feature of such "information borrowing" is that the practitioners can pick multiple values of the scale or discounting parameter to compare the analysis outcomes reflecting different prior beliefs of the "similarity" between the historical data and current data.

In this paper, we showed how the Bayesian analysis with power priors and non-informative priors can be used in modeling auto claims frequency for a new territory. By borrowing prior information from an existing territory that shares some similar characteristics such as climate, population demographics, or geography, we can develop a claim frequency model. Modeling claim severity with Bayesian GLM is also shown. We illustrated our approach on a small data set drawn from the motor third-party liability data set provided by a French insurance company. An immediate future work we would like to pursue is the joint Bayesian analysis of frequency and severity of claims. The joint analysis would allow us to borrow information between the two numerical features of the "new territory" and, hence, improve the analysis with a limited amount of information. The validation of our approach was also provided. We believe that our attempt to introduce Bayesian analysis with power priors will benefit many insurance companies as they enhance their current GLM pricing model and apply it in ratemaking of a new territory or an existing territory where the claims' experience is limited.

Acknowledgment

The authors are grateful for the comments and suggestions received by two anonymous Reviewers. We also appreciate the discussion with Professor John Zicarellie from Arizona State University whose suggestions further improved the clarity of this paper.

Supplementary Material

The SAS Code is available in Appendix B.

REFERENCES

[1] CASdatasets. URL: http://127.0.0.1:17326/library/CASdatasets/html/overview.html

[2] Charpentier, A. (2014). Computational Actuarial Science with R. CRC Press.

[3] Chen, M.H. and Ibrahim, J.G. (2006). "The relationship between the power prior and hierarchical models," *Bayesian Analysis* ,Vol 1, 551-574.

[4] Bermúdez, L. and Karlis, D. (2011). "Bayesian multivariate Poisson models for insurance ratemaking," *Insurance: Mathematics and Economics*, Vol 48, No. 2, 226-236.

[5] Boucher, J.P., Denuit, M. and Guillen, M. (2008). "Models of insurance claim counts with time dependence based on generalization of Poisson and negative binomial distributions," *Variance*, Vol 2, No 1, 135-162.

[6] Boucher, J.P., Denuit, M. and Guillen, M. (2009). "Number of Accidents or Number of Claims? An Approach with Zero-Inflated Poisson Models for Panel Data," *Journal of Risk and Insurance*, Vol *76, No* 4, 821-846.

[7] Brown, G.O. and Buckley, W.S. (2015). "Experience rating with Poisson mixtures," *Annals of Actuarial Science*, Vol 9, No 2, 304-321.

[8] Goldburd, M., Khare, A., and Tevet, D. (2016). "Generalized Linear Models for Insurance Rating", Casualty Actuarial Society, CAS Monographs Series, No 5.

[9] Ibrahim, J.G., and Chen, M.-H. (2000). "Power prior distributions for regression models," *Statistical Science*, Vol 15, 46-60.

[10] Kaas, R., Goovaerts, M., Dhaene, J. and Denuit, M. (2008). *Modern actuarial risk theory: using* R. Springer Science & Business Media.

[11] McCullagh, P., and J. A. Nelder, *Generalized Linear Models* (2nd edition), London: Chapman and Hall, 1989.

[12] Miljkovic, T. (2017). *Computational Actuarial Science with* R. Journal of Risk and Insurance 84(1): 267.

[13] Miljkovic T. and Fernández, D. (2018). *On Two Mixture-Based Clustering Approaches Used in Modeling an Insurance Portfolio.* Risks 6 (2). DOI: 10.3390/risks6020057.

Abbreviations and notations

GLM, generalized linear models

ROC, Receiver Operating Characteristic

REML, Restricted Maximum Likelihood

Biographies of the Authors

Dr. Jing Zhang is an Associate Professor in the Department of Statistics at Miami University. Her research focus is in Bayesian statistics, spatial analysis, statistical modeling for environmental and biological studies. Zhang holds an MBA and PhD in Statistics from University of Missouri.

Dr. Tatjana Miljkovic is an Assistant Professor and the Actuarial Science Adviser in the Department of Statistics at Miami University. Her research focus is in actuarial science and applied statistics areas. Prior to earning her PhD in statistics, she was employed by Unum Life Insurance Company (Corporate Actuarial), American National Property and Casualty Company (Corporate Actuarial), and Risk Management Solutions (Model Development). Miljkovic holds an MS Degree in Actuarial Science from University of Illinois, an MBA and PhD in Statistics from North Dakota State University.

Appendix A

Method	Frequentist Approach to Poisson Regression				Bayesian Poisson Regression with Weakly Informative Priors				Bayesian Poisson Regression with Power Priors			
Parameter	Estimate	SE	95% CI		Estimate	SE	95% CI		Estimate	SE	95% CI	
Intercept	-0.4754	0.2468	-0.9590	0.0083	0.8330	0.2513	0.3292	1.3164	0.7670	0.2027	0.3531	1.1498
GasRegular	-0.2582	0.1038	-0.4617	-0.0547	-0.2588	0.1047	-0.4676	-0.0553	-0.2329	0.0839	-0.3954	-0.0661
DriverAge(20, 26]	-0.3149	0.2502	-0.8052	0.1754	-0.4907	0.2519	-0.9811	0.0053	-0.4339	0.2044	-0.8302	-0.0287
DriverAge(26, 42]	-1.0424	0.2070	-1.4481	-0.6368	-1.1942	0.2091	-1.5883	-0.7662	-1.1039	0.1711	-1.4212	-0.7554
DriverAge(42, 74]	-0.8483	0.1946	-1.2297	-0.4670	-1.0565	0.1964	-1.4231	-0.6553	-1.0441	0.1613	-1.3456	0.7111
DriverAge(74, Inf]	-0.6587	0.2609	-1.1701	-0.1473	-0.9487	0.2635	-1.4637	-0.4322	-0.9194	0.2125	-1.3389	-0.5038
Density	0.1896	0.0393	0.1126	0.2667	0.1975	0.0392	0.1179	0.2718	0.1958	0.0328	0.1301	0.2579

Table-1: Comparisons of the Regression Results for Poisson Model

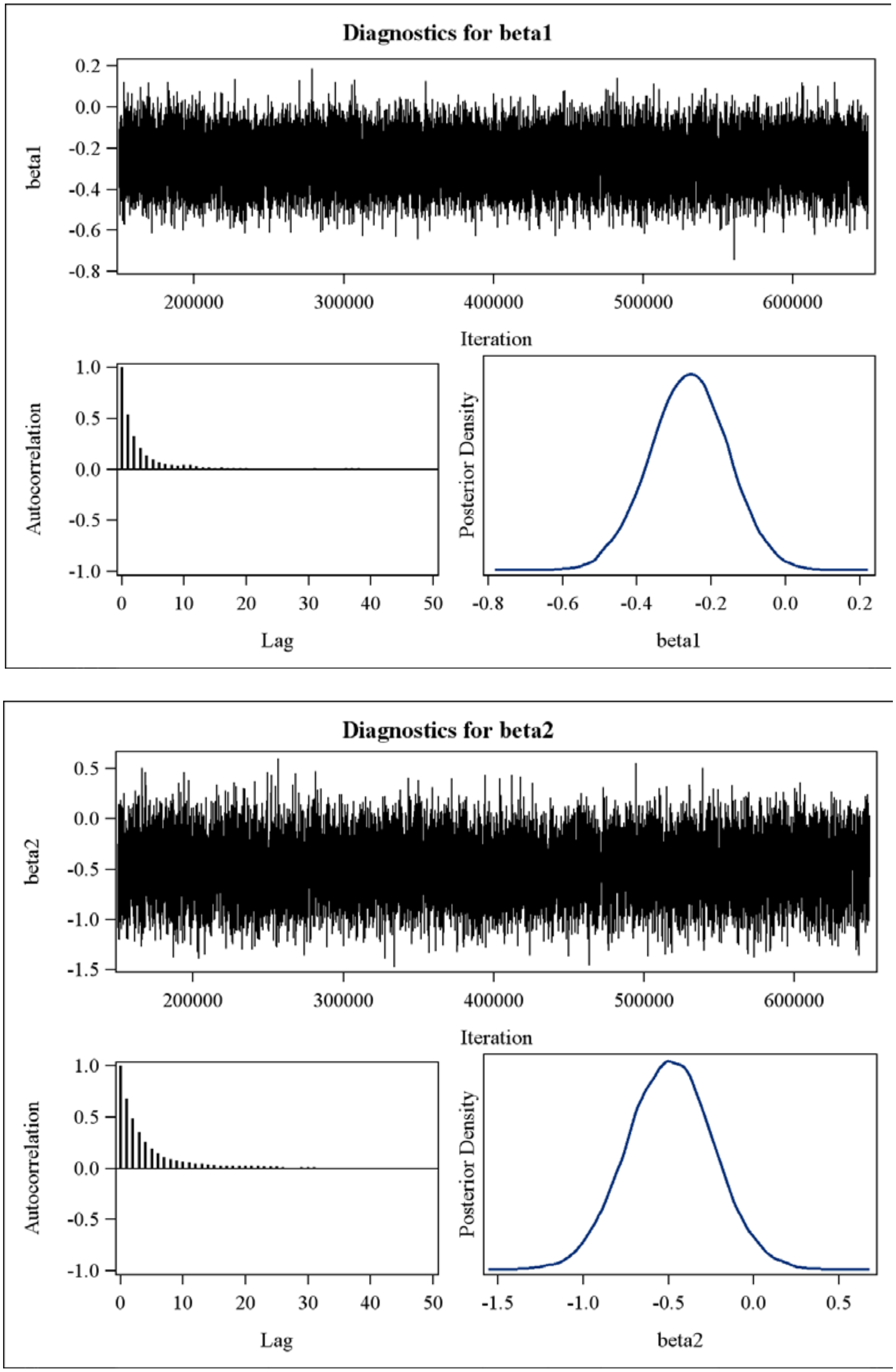
Table-2: Comparisons of the Regression Results for Gamma Model

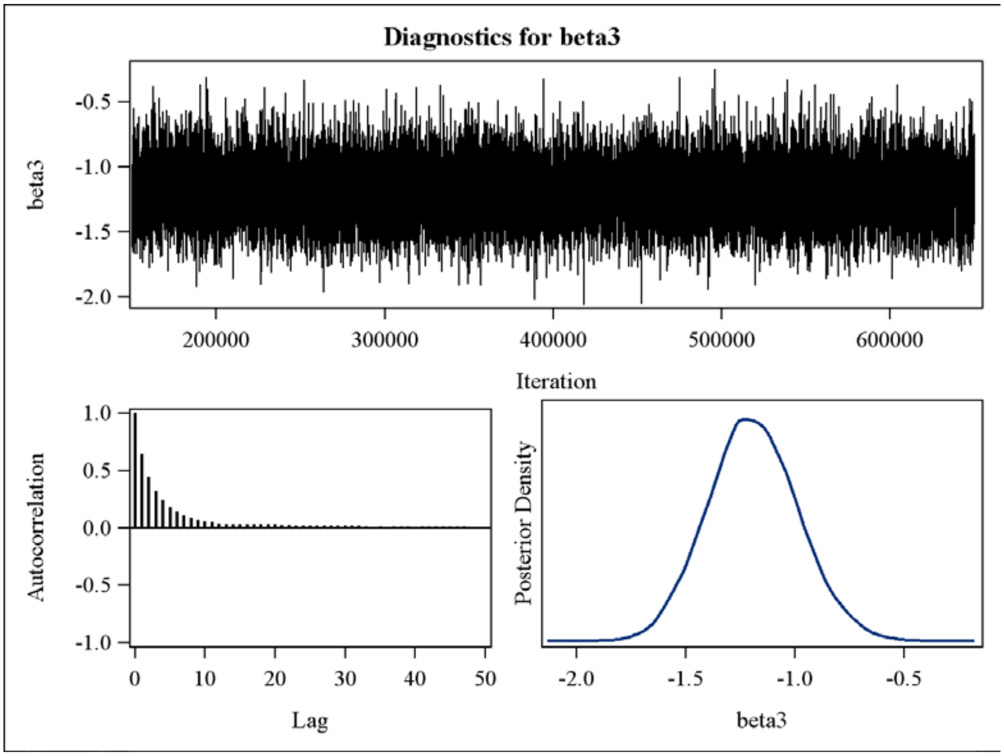
Method	Frequentist Approach to Gamma Regression				Bayesian Gamma Regression with Non-informative Priors			
Parameter	Estimate	SE	95% CI		Estimate	SE	95% CI	
Intercept	7.7746	0.1728	7.4359	8.1133	7.7869	0.1690	7.4650	8.1261
GasRegular	0.1125	0.0819	-0.0481	0.2730	0.1109	0.0818	-0.0513	0.2651
DriverAge(20,26]	-0.6454	0.2328	-1.1017	-0.1892	-0.6423	0.2320	-1.1074	-0.2037
DriverAge(26,42]	-0.4135	0.1773	-0.7611	-0.0659	-0.4218	0.1751	-0.7695	-0.0900
DriverAge(42,74]	0.2735	0.1727	-0.0650	0.6120	0.2652	0.1700	-0.0882	0.5741
DriverAge(74,Inf]	0.2555	0.2302	-0.1957	0.7067	0.2587	0.2308	-0.1743	0.7226
Density	0.0453	0.0426	-0.0382	0.1287	0.0474	0.0430	-0.0365	0.1327

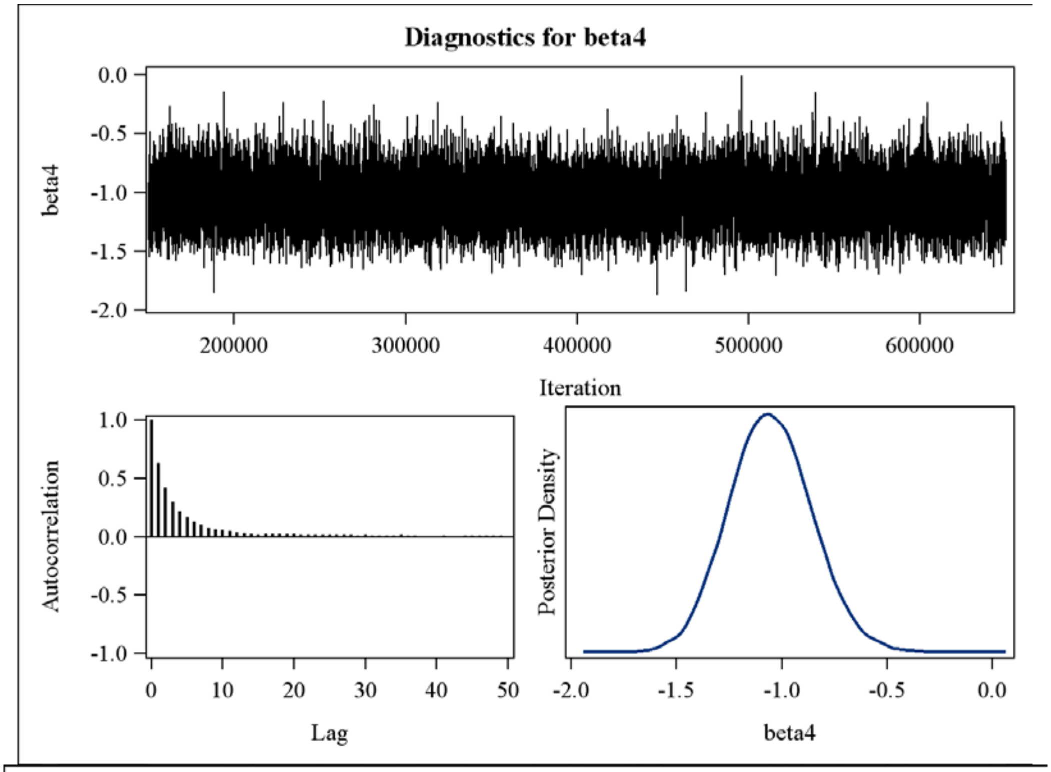
Table-3: Validation analysis

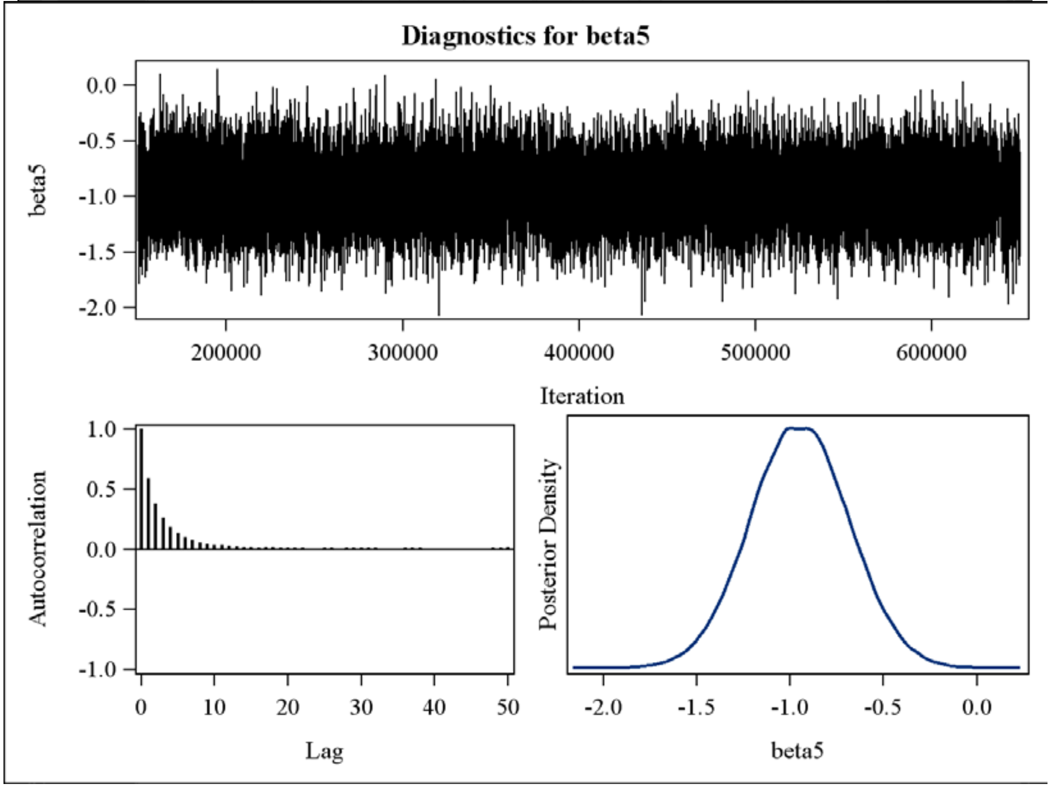
Sample Type	Training Set				Holdout Set				External Prior Set			
Parameter	Estimate	SE	95% CI		Estimate	SE	95% CI		Estimate	SE	95% CI	
Intercept	-0.4754	0.2468	-0.9590	0.0083	0.1793	0.8906	-1.5662	1.9248	-0.6566	0.2423	-1.1315	-0.1817
GasRegular	-0.2582	0.1038	-0.4617	-0.0547	-0.6379	0.3898	-1.4020	0.1261	-0.1784	0.1025	-0.3793	0.0225
DriverAge(20, 26]	-0.3149	0.2502	-0.8052	0.1754	-0.6266	0.8770	-2.3454	1.0923	-0.2215	0.2427	-0.6972	0.2542
DriverAge(26, 42]	-1.0424	0.2070	-1.4481	-0.6368	-1.2490	0.7849	-2.7875	0.2894	-0.8161	0.2024	-1.2129	-0.4193
DriverAge(42, 74]	-0.8483	0.1946	-1.2297	-0.4670	-0.8827	0.7406	-2.3343	0.5689	-0.8489	0.1955	-1.2321	-0.4658
DriverAge(74, Inf]	-0.6587	0.2609	-1.1701	-0.1473	-1.8648	1.2715	-4.3569	0.6272	-0.6206	0.2609	-1.1318	-0.1093
Density	0.1896	0.0393	0.1126	0.2667	0.2220	0.1620	-0.0955	0.5395	0.1744	0.0422	0.0916	0.2571

Figure 4. History plots, ACF plots and density curves of the model parameters in the Bayesian Poisson regression model with weakly informative priors.




Diagnostics for beta3
beta3
-0.5
-1.0
-1.5
-2.0
200000
300000
400000
500000
600000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-2.0
-1.5
-1.0
-0.5
beta3


Diagnostics for beta4
beta4
0.0
-0.5
-1.0
-1.5
-2.0
200000
300000
400000
500000
600000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-2.0
-1.5
-1.0
-0.5
0.0
beta4


Diagnostics for beta5
beta5
0.0
-0.5
-1.0
-1.5
-2.0
200000
300000
400000
500000
600000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-2.0
-1.5
-1.0
-0.5
0.0
beta5

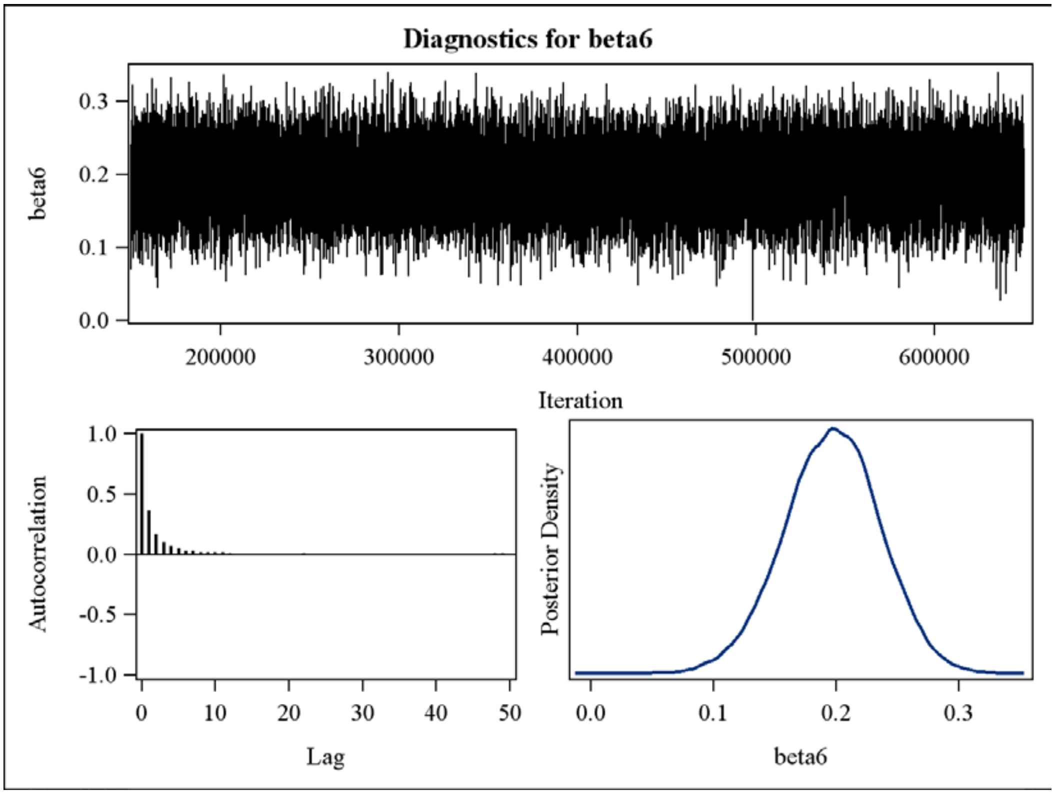
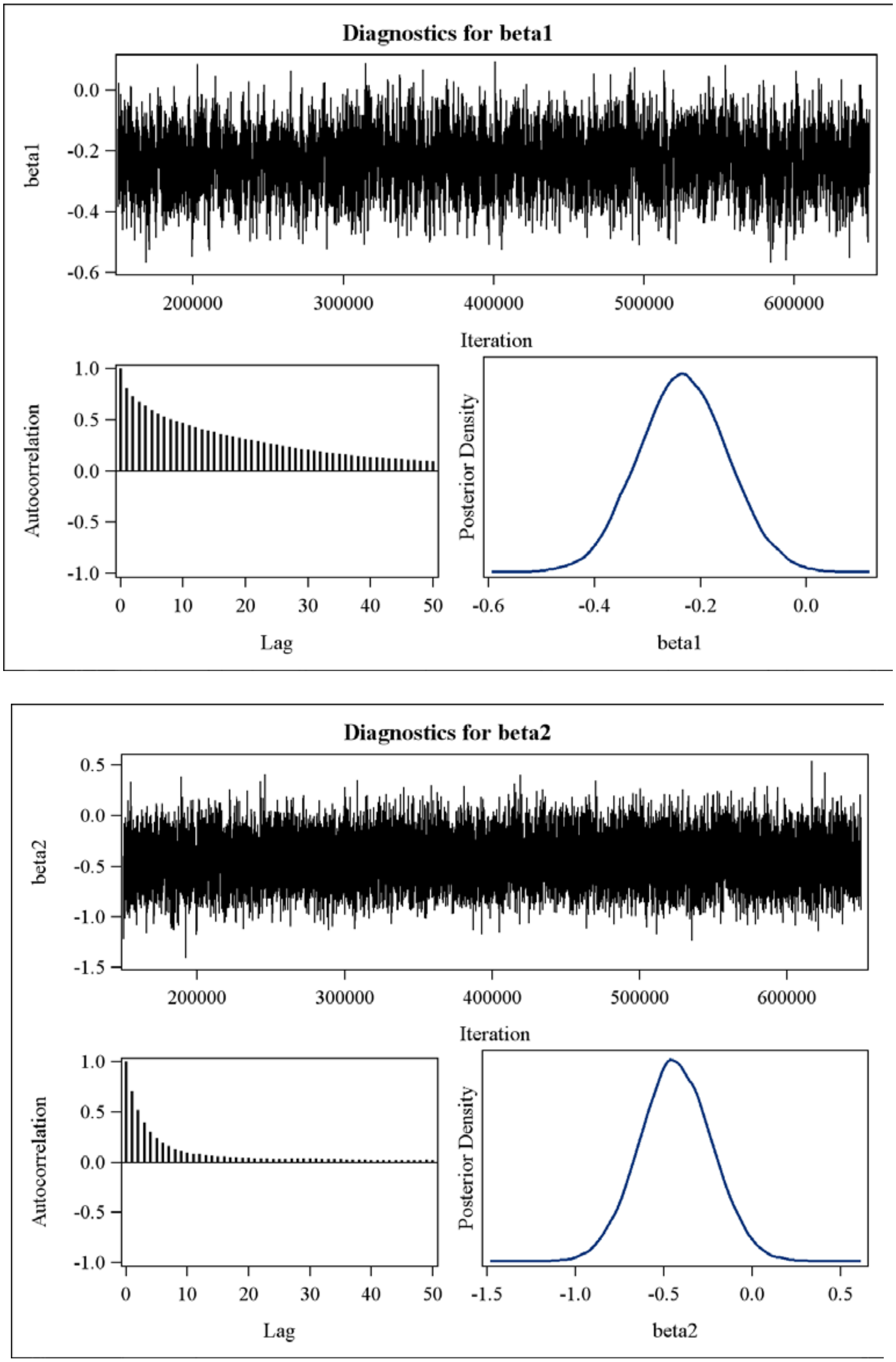
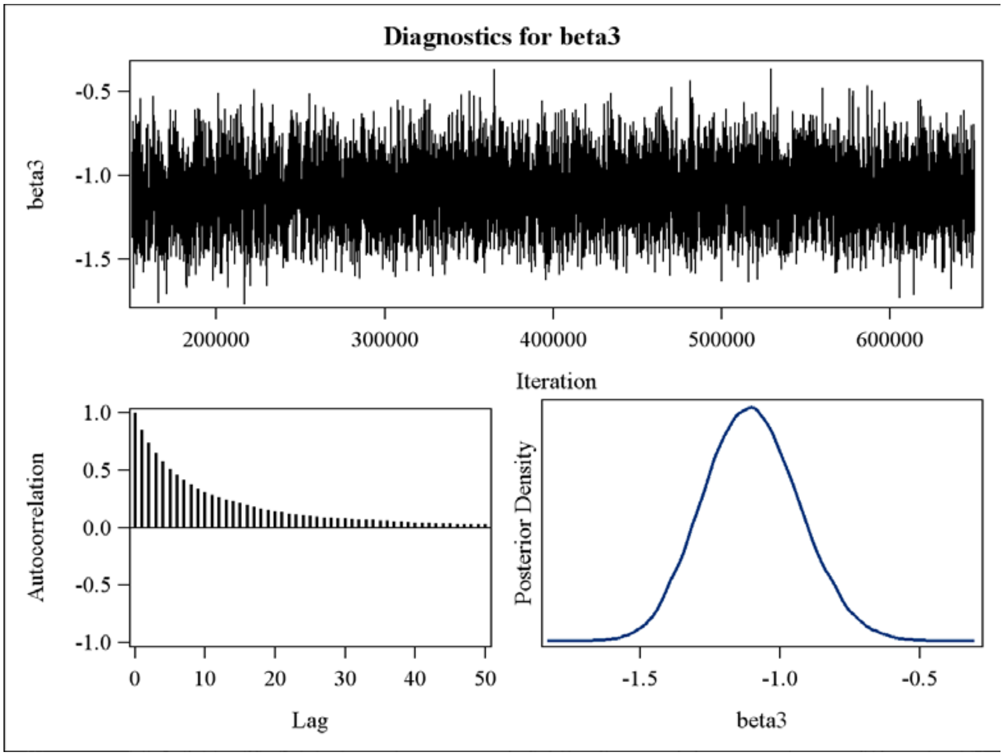
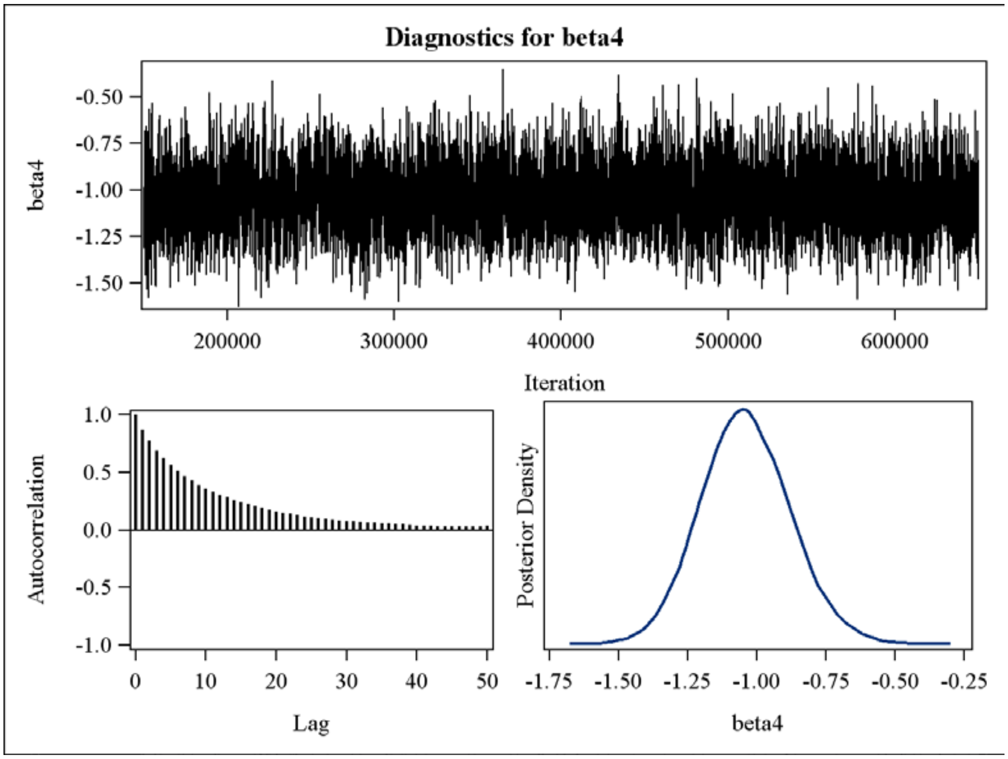
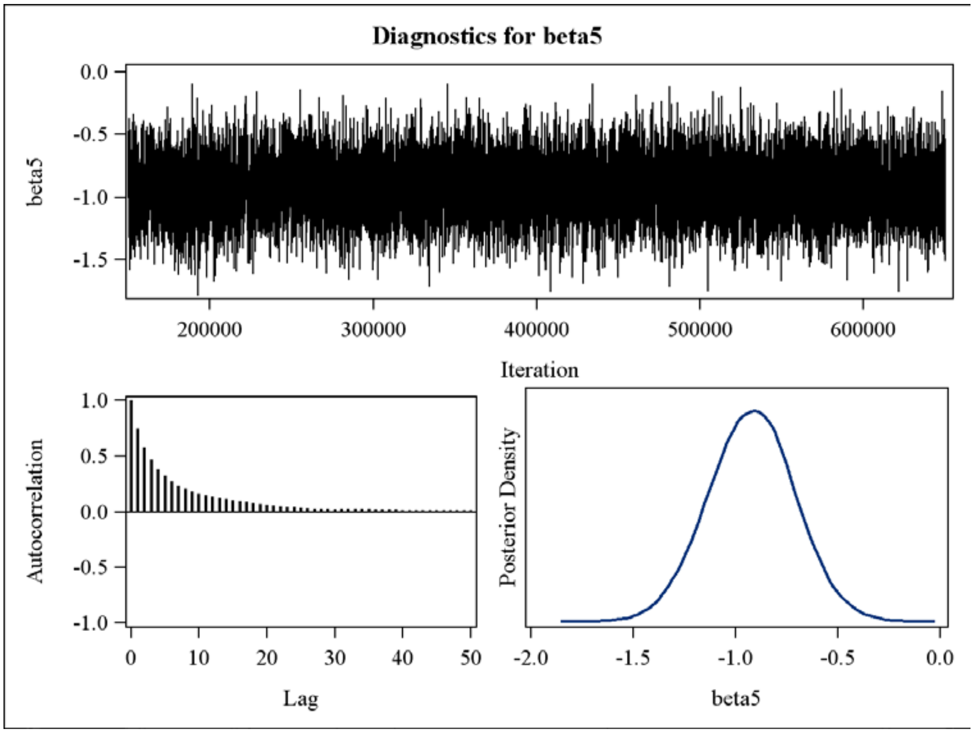

Diagnostics for beta6
beta6
0.0
0.1
0.2
0.3
200000
300000
400000
500000
600000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
0.0
0.1
0.2
0.3
beta6

Figure 5. History plots, ACF plots and density curves of the model parameters in the Bayesian Poisson regression model with power priors.




Diagnostics for beta3
beta3
-0.5
-1.0
-1.5
200000
300000
400000
500000
600000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-1.5
-1.0
-0.5
beta3


Diagnostics for beta4
beta4
-0.50
-0.75
-1.00
-1.25
-1.50
200000
300000
400000
500000
600000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-1.75
-1.50
-1.25
-1.00
-0.75
-0.50
-0.25
beta4


Diagnostics for beta5
beta5
0.0
-0.5
-1.0
-1.5
200000
300000
400000
500000
600000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-2.0
-1.5
-1.0
-0.5
0.0
beta5

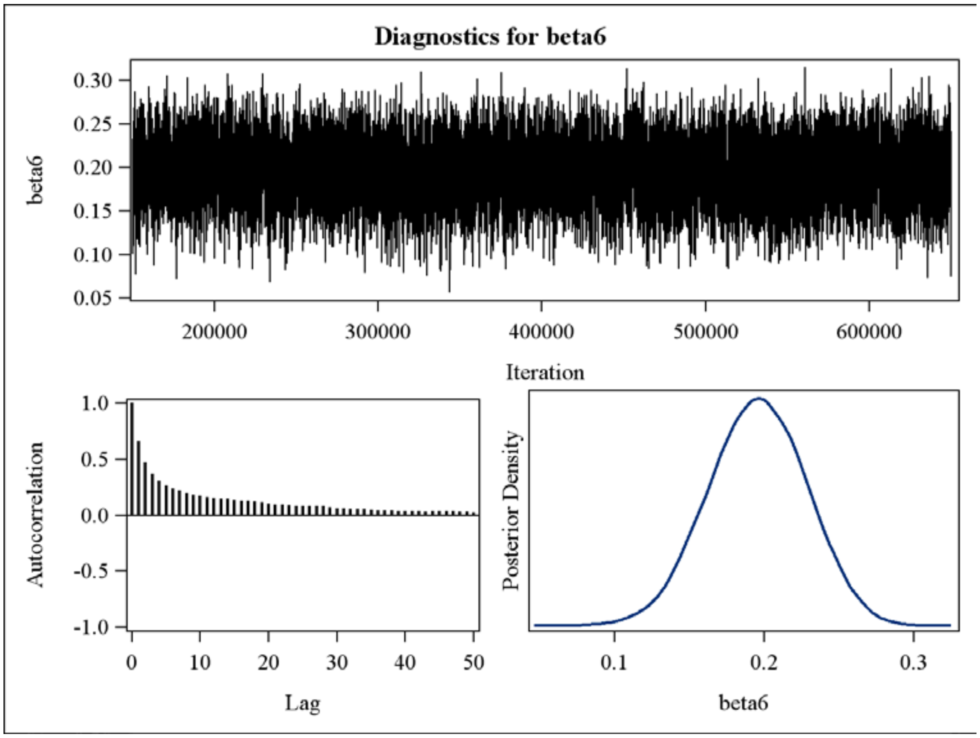
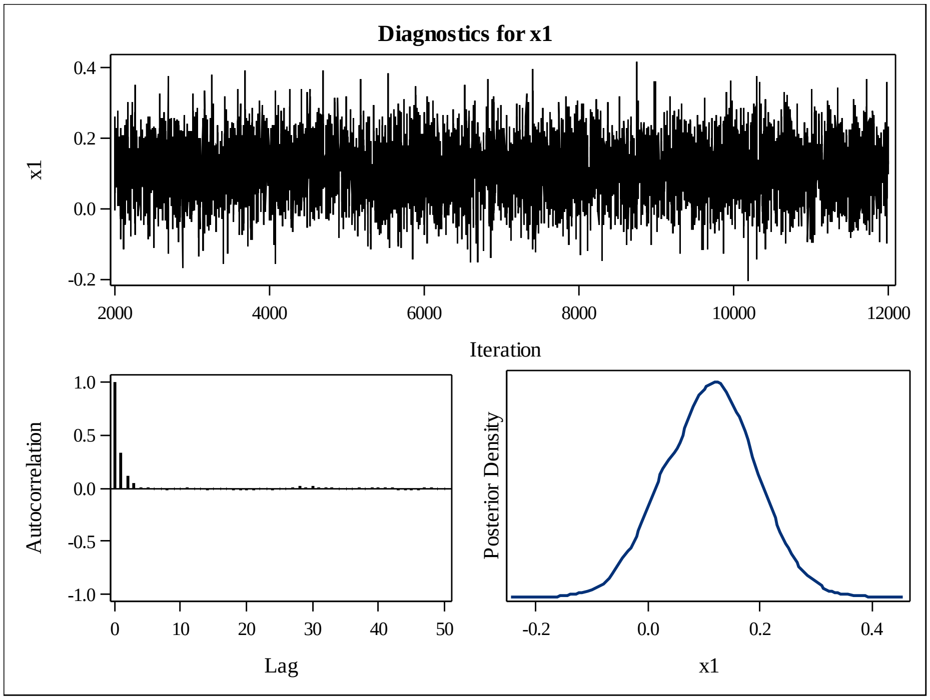
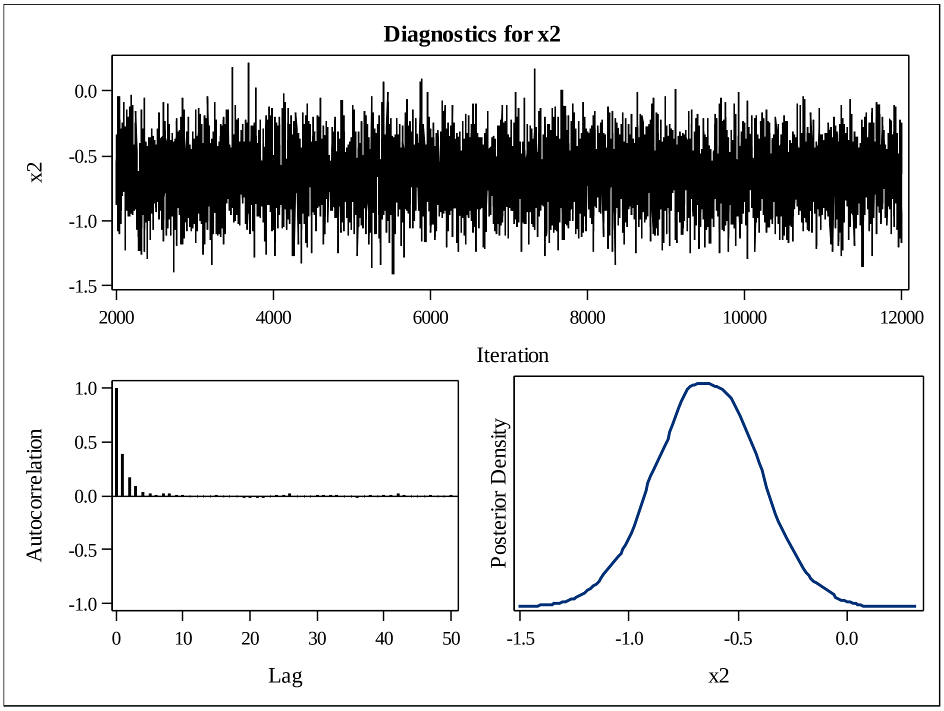
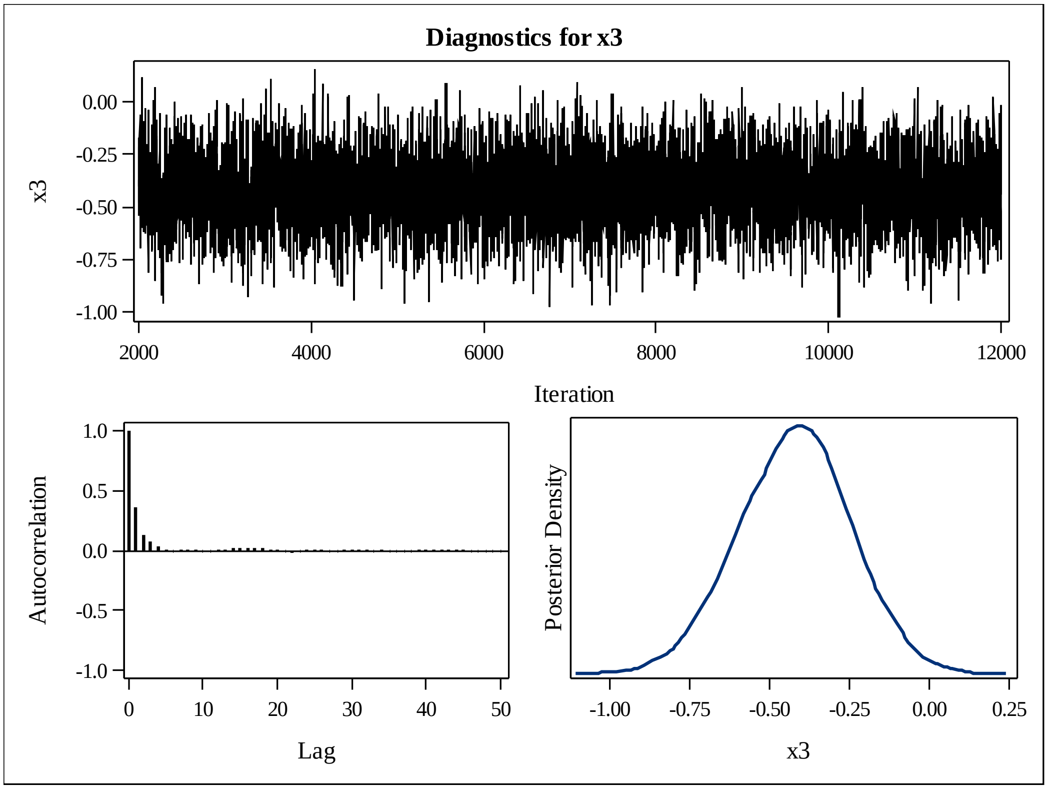
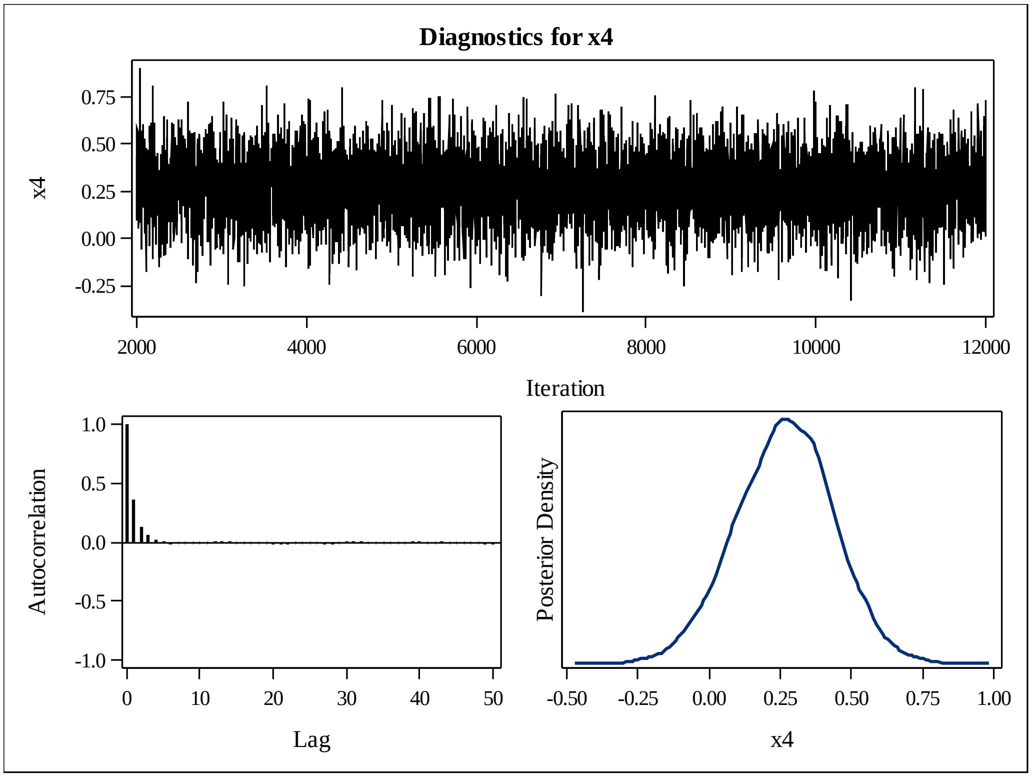

Diagnostics for beta6
beta6
0.30
0.25
0.20
0.15
0.10
0.05
200000
300000
400000
500000
600000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
0.1
0.2
0.3
beta6

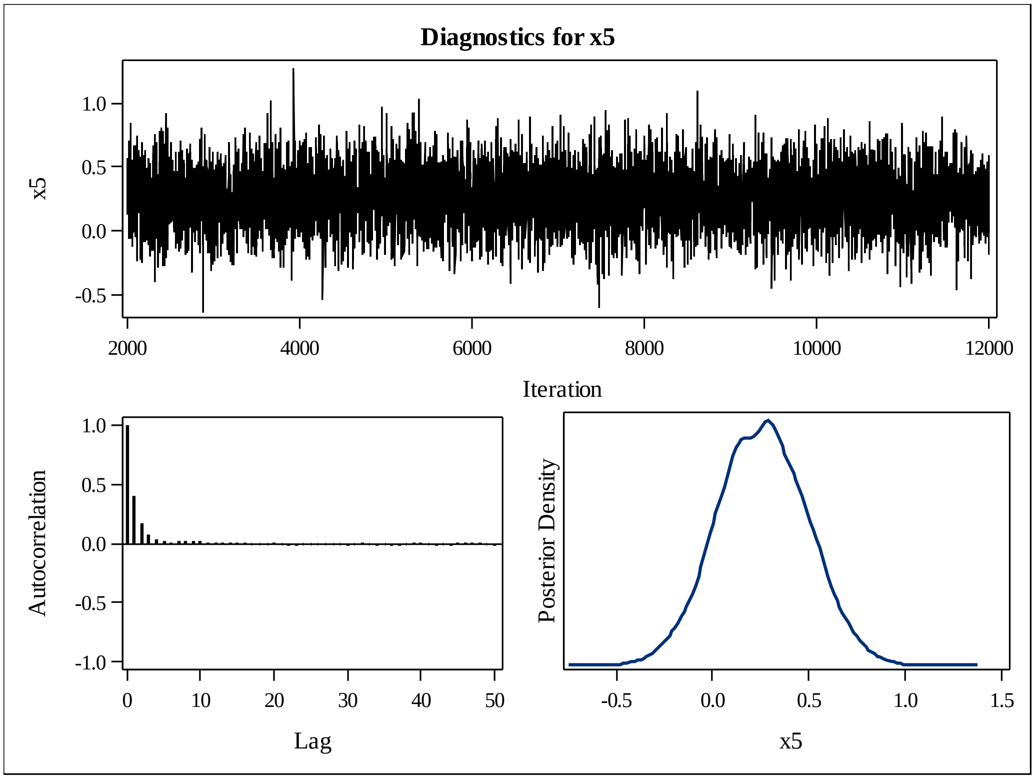
Figure 6. History plots, ACF plots and density curves of the model parameters in the Bayesian gamma regression model.

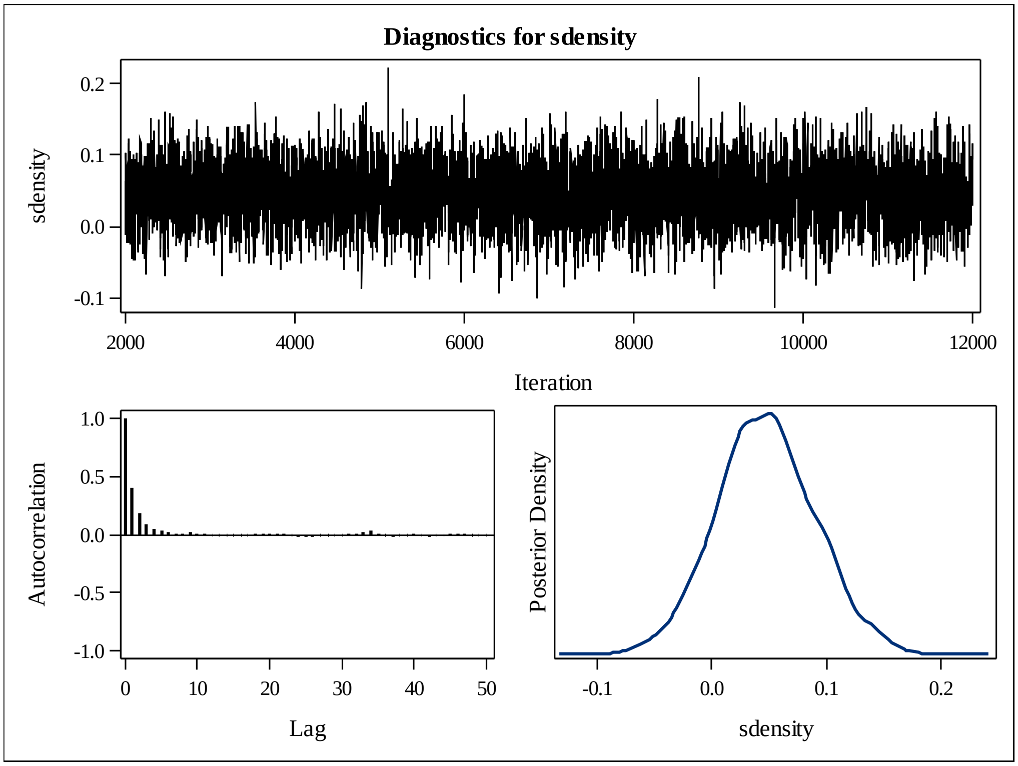





Diagnostics for x3
x3
0.00
-0.25
-0.50
-0.75
-1.00
2000
4000
6000
8000
10000
12000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-1.00
-0.75
-0.50
-0.25
0.00
0.25
x3


Diagnostics for x4
x4
0.75
0.50
0.25
0.00
-0.25
2000
4000
6000
8000
10000
12000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-0.50
-0.25
0.00
0.25
0.50
0.75
1.00
x4


Diagnostics for x5
x5
1.0
0.5
0.0
-0.5
2000
4000
6000
8000
10000
12000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-0.5
0.0
0.5
1.0
1.5
x5


Diagnostics for sdensity
sdensity
0.2
0.1
0.0
-0.1
2000
4000
6000
8000
10000
12000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
-0.1
0.0
0.1
0.2
sdensity

Appendix B

SAS code for the analysis.
Part 1:

```
/*Poisson regression with weakly-informative priors*/
/*
ClaimNb: response variable, number of claims occurred during a
given time period in the region for a customer;
logoffset: logarithm of the exposure (duration of the policy),
offset variable we used in the Poisson regression;
Gas: indicator variable, Gas=1 if the car insured uses regular
gas;
x2: indicator variable, x2=1 if the insured driver is older than
22 but is 26 or younger;
x3: indicator variable, x3=1 if the insured driver is older than
26 but is 42 or younger;
x4: indicator variable, x4=1 if the insured driver is older than
42 but is 74 or younger;
x5: indicator variable, x5=1 if the insured driver is older than
74;
Density: population density of the region;
alpha: intercept;
beta1-beta6: regression coefficients associated with Gas, X2-X5
and Density;

Priors used for alpha: alpha ~ N(mua,sda^2), with higher level
priors mua~N(0, 10) and sda~Uniform(0,5).

Priors used for alpha and all the betai's: N(mubi,sdbi), i=1,...,6,
with higher level priors mubi~N(0,4) and sdbi~Uniform(0,1).
*/
```

```
proc mcmc data=datasetname seed=1181 nmc=500000 nbi=150000 thin=10
propcov=quanew monitor =(_parms_ ) outpost=out1000prior2;
ods select Parameters PostSummaries PostIntervals tadpanel;
parms alpha 0 beta1 0 beta2 0 beta3 0 beta4 0 beta5 0 beta6 0;
parms mua 0 mub1 0 mub2 0 mub3 0 mub4 0 mub5 0 mub6 0;
parms sda 0.5 sdb1 0.5 sdb2 0.5 sdb3 0.5 sdb4 0.5 sdb5 0.5 sdb6 0.5;
prior alpha ~ normal(mua, var=sda**2);
prior beta1 ~ normal(mub1, var=sdb1**2);
prior beta2 ~ normal(mub2, var=sdb2**2);
prior beta3 ~ normal(mub3, var=sdb3**2);
prior beta4 ~ normal(mub4, var=sdb4**2);
prior beta5 ~ normal(mub5, var=sdb5**2);
prior beta6 ~ normal(mub6, var=sdb6**2);
prior mua ~ normal(0, var=10);
prior mub: ~ normal(0, var=4);
prior sda ~ uniform(0,5);
prior sdb: ~ uniform(0,1);

mu = exp(logoffset + alpha + beta1*Gas +
beta2*x2+beta3*x3+beta4*x4+beta5*x5+beta6*Density);
model ClaimNb ~ poisson(mu);
run;
```

Part 2:

```
/*Poisson regression with Power prior, a0 fixed*/

/*
ClaimNb: response variable, number of claims occurred during a
given time period in the region for a customer;
logoffset: logarithm of the exposure (duration of the policy),
offset variable we used in the Poisson regression;
Gas: indicator variable, Gas=1 if the car insured uses regular
gas;
x2: indicator variable, x2=1 if the insured driver is older than
22 but is 26 or younger;
x3: indicator variable, x3=1 if the insured driver is older than
26 but is 42 or younger;
x4: indicator variable, x4=1 if the insured driver is older than
42 but is 74 or younger;
x5: indicator variable, x5=1 if the insured driver is older than
74;
Density: normalized population density of the region;

alpha: intercept;
beta1-beta6: regression coefficients associated with Gas, X2-X5
and Density;

Initial Priors used for alpha: alpha ~ N(mua,sda^2), with higher
level priors mua~N(0, 10) and sda~Uniform(0,5).

Initial Priors used for alpha and all the betai's: N(mubi,sdbi),
i=1,...,6, with higher level priors mubi~N(0,4) and
sdbi~Uniform(0,1).

Power prior is used here with fixed power a0=0.5.
*/
proc mcmc data=datasetname seed=1181 nmc=500000 nbi=150000 thin=10
```

```
propcov=quanew monitor =(_parms_ ) outpost=out1000power50;
ods select Parameters PostSummaries PostIntervals tadpanel;
parms alpha 0 beta1 0 beta2 0 beta3 0 beta4 0 beta5 0 beta6 0;
parms mua 0 mub1 0 mub2 0 mub3 0 mub4 0 mub5 0 mub6 0;
parms sda 0.5 sdb1 0.5 sdb2 0.5 sdb3 0.5 sdb4 0.5 sdb5 0.5 sdb6 0.5;
prior alpha ~ normal(mua, var=sda**2);
prior beta1 ~ normal(mub1, var=sdb1**2);
prior beta2 ~ normal(mub2, var=sdb2**2);
prior beta3 ~ normal(mub3, var=sdb3**2);
prior beta4 ~ normal(mub4, var=sdb4**2);
prior beta5 ~ normal(mub5, var=sdb5**2);
prior beta6 ~ normal(mub6, var=sdb6**2);
prior mua ~ normal(0, var=10);
prior mub: ~ normal(0, var=4);
prior sda ~ uniform(0,5);
prior sdb: ~ uniform(0,1);
begincnst;
a0=0.5;
endcnst;
mu = exp(logoffset + alpha + beta1*Gas +
beta2*x2+beta3*x3+beta4*x4+beta5*x5+beta6*Density);
llike=logpdf('poisson',ClaimNb,mu);
if (city='old') then llike=a0*llike;
model general(llike);
run;
```

 Part 3:

```
/*Gamma regression with noninformative prior for severity
analysis*/

/*
AggClaimAmount: response variable, severity of claims;
ClaimNb: number of claims occurred during a given time period in
the region for a customer, used as the exponential family
dispersion parameter weight for each observation;
X1: indicator variable, Gas=1 if the car insured uses regular gas;
x2: indicator variable, x2=1 if the insured driver is older than
22 but is 26 or younger;
x3: indicator variable, x3=1 if the insured driver is older than
26 but is 42 or younger;
x4: indicator variable, x4=1 if the insured driver is older than
42 but is 74 or younger;
x5: indicator variable, x5=1 if the insured driver is older than
74;
sdensity: normalized population density of the region;
Default uniform priors are used for all regression coefficients;
Default INV-Gamma(0.001, 0.001) used for the rate parameter (see
paper).
*/


proc genmod data= datasetname;
class Gas;
Weight ClaimNb;
model AggClaimAmount = x1 x2 x3 x4 x5 sdensity/ dist=gamma
                             link=log;
bayes seed=4 outpost=postgamma diagnostics=all summary=all;
run;
```

Meteorology for Actuaries – Part 2
Climate and the El Niño Southern Oscillation

Gwendolyn Anderson, ACAS, MAAA

The intertwining threats of climate change and catastrophe challenge society's ability to interpret shocks like recent hurricanes and wildfires. New capabilities have arisen in the recently expanded power of home computers which can now process vast databases; and in shared tools, such as R programming which offers calculation tools in combination with palatable "visual analysis" through plots and maps. Utilizing these technologies, this paper serves as a reference guide to weather analysis as it pertains to climate, and as regional climates relate to loss. A high level of detail in daily station records allows matching of specific weather measurements to losses in both time and location, lending ability to identify thresholds, durations, and combined forces leading to loss; further, changes in data or data quality can then be distinguished from shifts in climate. Physical explanations provide essential directions to begin exploration, focusing on an example of the phases of El Niño Southern Oscillation (ENSO) by which climate varies throughout the globe naturally, not only in extremes. The venture to discover climate's effect on losses becomes less daunting through pre-written modifiable code, sources for ENSO indices and other meaningful inputs, and a useful collection of tables and visual references.

Availability: https://cran.r-project.org/
ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/

Keywords: Climate change, weather analysis, El Niño Southern Oscillation (ENSO), R programming, maps

Abbreviations:

ENSO	El Niño Southern Oscillation
GHCN-D	Global Historical Climatology Network – Daily ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/readme.txt
NOAA	National Oceanic & Atmospheric Administration
SST	Sea Surface Temperature

GHCN Weather Elements

`PRCP`	Precipitation
`SNOW`	Snowfall
`SNWD`	Snow depth
`TMAX`	Maximum temperature
`TMIN`	Minimum temperature

`WIND*` elements are coded to include:

`AWND`	Average daily wind speed
`WSF1`	Fastest 1-minute wind speed
`WSF2`	Fastest 2-minute wind speed
`WSF5`	Fastest 5-second wind speed
`WSFG`	Peak gust wind speed
`WSFI`	Highest instantaneous wind speed
`WSFM`	Fastest mile wind speed

(* `WIND` is *not* an element abbreviation of GHCN-D.)

1. INTRODUCTION

Part One provided a basic backdrop of maps that can be instantly plotted in R language. Weather and one's own loss values may now be added to these backdrops. Part Two introduces daily meteorological data, publicly available through the Global Historical Climatology Network – Daily (GHCN-D). These vast datasets offer the level of detail suited to matching with weather-related losses in both time and location, easily accessed by R code with 8GB memory, a recent standard for most home computers. These combined advancements – memory, language, and data – expand the potential for exploring not only weather events but overall shifts of climate. This paper provides code, input sources and references, along with physical explanations of the weather phenomenon. Climate cycles are illustrated through an example of the El Niño Southern Oscillation. Many maps and plots in this paper are produced in R language from modifiable code provided in the appendix.

1.1 Research Context

As extreme weather events devastate North America, continually breaking records of a recent past, concerns widen over what seem to be pronounced changes in climate: is the potential for change understood well enough to simply prepare for the next storm? In absence of human industry, climate already changes naturally, with a myriad of interactions from diverse sources on multiple scales. Adding layers of complexity is the growing range of human activities that appear to impact climate systems, all while human skills and technologies advance in sync with nature's destruction. Portentous storms assert the need to utilize modern technologies to a timely advantage, to place state-of-the-art tools in reach to those with both common and uncommon skills.

1.1.1 Record Storm Losses

The costliest storms in United States history, those producing damages of $1 billion or more, are plotted below chronologically in actual unadjusted costs. A notable escalation of events occurred in the last three decades, disrupting the scale of catastrophic loss. Hurricane Hugo took a destructive inland course in 1989, followed in 1992 by Andrew which more than doubled record cost in three short years. Andrew led to insurer insolvencies, sending shock and a wake-up call through the industry. Professional leaders then turned to catastrophe modeling for answers, simulating the physical process of hurricane activity within trade secret models. This move proved effective in preparing financially for the spate of mega storms to follow.

By 2005, Katrina seemed to break the all the rules, striking levees and storm walls which had not been properly engineered to prevent vast flooding of the low-lying New Orleans area. Failed planning, it seemed, had quadrupled the former record set by Andrew.

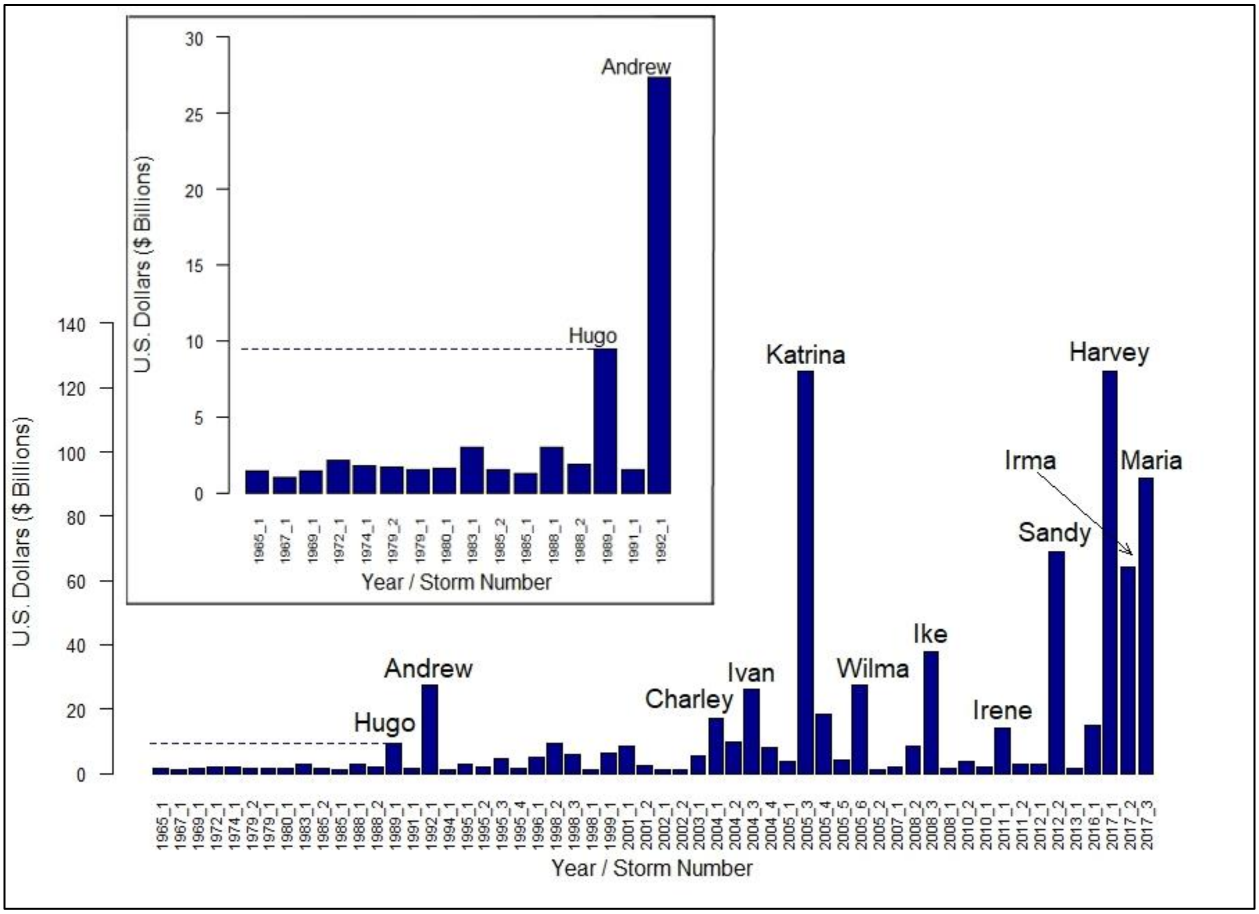


Figure 1. Costliest Atlantic Hurricanes - Katrina damages were a multiple of preceding "mega storms" due to failed engineering. In absence of adjustments, more recent storms seem to rescale catastrophes since the post-Andrew era. **Inset** - Andrew damages were unprecedented and took insurers by surprise. (All storms exceeding U.S. $1B in actual, unadjusted damages.)

Yet in 2012, Sandy struck the most densely populated area of the United States, unusually far to the north and late in season, again destroying property at multiple times the scope of Andrew. Five years later, the combined severity of three major land falling hurricanes in 2017 is unprecedented for a single season, with no poor engineering to blame. It is clear from these pictures that the costliest storms in all of history are also the most recent – imagine the shape had the Gulf Coast been protected against Katrina. These trends would appear to belie randomness and raise new questions surrounding severe weather and climate. In response to public concern, the role and sources of climate change might now be approached across broader disciplines.

1.1.2 A New Kind of Global Warming

Actuaries, modelers, economists and scientists alike are inclined to bring the historical loss record in line with present day conditions. Such adjustments include consideration for not only dollar inflation, but construction upgrades, migrations of population to coastlines, changes in levels of wealth near ocean fronts; and in the case of an insurance loss history, any changes in coverages or generosity of settlement. These financial insights alter the picture entirely: by various estimates of 'normalized' storm damages spanning over a century, the outcome appears a random process. Considering storms since 1900, the ICAT Damage Estimator [www.icatdamageestimator.com] ranks Katrina only as the fourth most damaging storm through 2012, and Sandy as eighth.

Hidden within the appearance of randomness is another process, recognized mainly in the scientific community, that might invite further refinements to views of weather-related losses. Somewhere in between the view of escalating catastrophes and the view of random losses, lies the natural force of *cyclical* climate change. The El Niño Southern Oscillation (ENSO) exerts major influence on the strength and timing of Atlantic hurricanes. Such cycles exist in absence of any human contribution to the atmosphere, and are irregular, reflecting a certain randomness of nature occurring in phases. These cycles may also be prone to change and may in themselves be subject to influences. How should climate be regarded if natural cycles might differ in frequency, duration or amplitude over future decades compared to the past century?

The same weather patterns or phases that influence severe weather events can be discerned more plainly in common weather elements like rainfall and temperature. These elements might attract less media attention than hurricanes, but will provide a far less volatile example of natural climate cycles. As a base illustration, Florida rainfall will be compared to ENSO indices in winter months, outside of the hurricane season.

The actuary, whose forté is prediction from limited data, might benefit from stepping into the shoes of the scientist, and might even tighten a few laces to fit a loss perspective. The weather data history is fraught with missing records and changes of stations whose measurements depend on elevation and surrounding conditions. If inflation, population, construction, wealth and coverage were not enough, changes in record keeping could also be mistaken for 'climate change.' The insurance industry may wish to weigh in on weather data collection now, to better account for climate shifts as they arrive.

A new trend in global warming may be in sight: a warming up to a cooperation in use of resources, from shared tools to shared understanding. Perhaps this trend may lead from strong varied opinions toward the exploration of facts and figures reinforced by the science that explains the physical phenomenon of weather.

1.2 Objective

Since research on weather and climate comes primarily from outside of the insurance sector, little focus is placed on loss estimates. Within the insurance industry, most research on these topics remains proprietary, limiting the public's grasp of the situation and limiting participation by those who might strengthen the climate conversation. This paper seeks to remove limitations to analysis. At most, the boundlessness of relationships to be explored might be recognized. At least, a highly detailed resource of element measurements should illuminate the sparseness of record available by which to identify climate shifts.

This paper provides tools and references to accompany the vast daily station data of the Global Historical Climatology Network (GHCN), along with an understanding of the physical processes of weather as insight to the analysis of weather peril losses. A framework is provided through code and useful references, with an example of the El Niño Southern Oscillation and Florida winter rainfall. Broad paths may be explored through this data set, whether the direction one wishes to pursue is global or focused within a unique region.

1.3 Outline

Background and Methods, Section 2, suggests methods for matching daily weather data sets to losses, through focus on damageability thresholds, durations, and interactions of weather elements that lead to loss. Beyond a programming method, a background in the physical phenomenon of weather guides interpretation of the data set and provides a basis for analysis. The description begins with the source of weather: the heat of the equator. Next, the motion of weather enters through the atmospheric circulation by which the heat is redistributed on earth. This leads to the core phenomenon to be covered, the El Niño Southern Oscillation (ENSO), with its pronounced influence on weather patterns in parts the globe distant from its origination around the equatorial Pacific. Sources are provided for indices that measure different oceanic regions of the ENSO phase by Sea Surface Temperature (SST), pressure, and other attributes. The time scale of 'climate' is differentiated from that of 'weather,' and cycles are recognized as a climate determinant. The short history of meteorological records gives insight into the sparseness of measurement available to compare climate over time, in spite of large data sets available today. The actuary's unique capabilities where data is lacking could be constructive contributors in the climate arena.

Results, Section 3, presents example findings from code output, including United States maps of station locations; and choropleths, anomalies color-coded by state. Some summaries of missing records and data changes are given by year. Florida winter rainfall is shown to correlate well to some ENSO indices and not to others.

Code to process the GHCN-D data is provided in the appendix. The code is intended for modification to the level desired.

2. BACKGROUND AND METHODS

The relationship between weather perils and insurance losses may be explored by linking a history of loss and exposures to the meteorological data. An approach is desired that would isolate the types of weather events leading to loss. Some background in the physical process of weather provide necessary insights to the analysis, bringing awareness of the natural climate cycles of the El Niño Southern Oscillation. Data quality and completeness require attention so that data changes not be mistaken for 'climate change.'

2.1 Thresholds

A high level of detail in daily station records allows matching of specific weather measurements to losses in both time and location. This detail lends the ability to establish thresholds at which losses are likely to occur. Thresholds tend to represent physical phenomena, such as zero degrees Celsius at which water freezes, or wind speeds that topple trees.

Durations of extreme weather are also relevant, and can be tracked daily up to the time of loss, such as low levels of precipitation eventually leading to crop loss. Combined forces may lead to damages, such as drought accompanied by high temperatures. Damaging interaction of weather elements may be intertemporal, such as drought-inflicted regions becoming susceptible to fires or mudslides with higher temperatures or rainfall, respectively. Thresholds should be expected to vary by region, for instance, Seattle with its immense drainage capacity may withstand multiples the rainfall of flood-prone Charleston.

Thresholds and durations cannot be extracted from monthly summaries, and loss events cannot be pinpointed in data sets that have been gridded in rectangular areas encompassing multiple stations. A maximum monthly temperature or average monthly temperature is not useful. Summaries that count threshold values can be created from daily data while retaining the source detail. Care must be taken to adjust for various changes to daily record keeping over time.

For the purpose of measuring climate change, standard deviation anomalies from a selected base period average serve as straightforward and meaningful measures. The anomaly will usually be calculated for a summary period, such as a month or year, compared to some longer base of 30 Januaries or 30 full years, for instance. These figures give an intuitive sense of fluctuation across time with appropriate scaling for the selected region; a large anomaly of rainfall in the desert will represent a small quantity in comparison to the same anomaly in the tropics. In absence of climate change, anomalies will take values spread about zero according to the stable underlying distribution of the element. Relative values like temperatures will be normally distributed while quantities like precipitation, bounded below by zero, will be skew.

Some regions suffer no loss from large deviation weather events while others regions hover at the edge of the climate extremes where disasters occur. Anomalous weather events could impact loss if present climate extremes of a region are close to loss thresholds. Attention should be given to cycles and shifts of climate in regions where loss thresholds have been crossed or where near-threshold weather patterns can be identified. The distribution of the weather element could be tracked over time or compared against the base period.

One familiar threshold guide is the Saffir-Simpson scale, which assigns a level of damage to hurricane categories by wind speed. The types of damages will vary by region and by the types of buildings in the region. The same damageability scales would clearly not apply in a country with building standards inferior to those of the United States.

Table 1. Saffir-Simpson Hurricane Wind Scale

Category	Sustained Winds	Damages
1	74-95 mph	Very dangerous winds will produce some damage
2	96-110 mph	Extremely dangerous winds will cause extensive damage
3	111-129 mph	Devastating damage will occur
4	130-156 mph	Catastrophic damage will occur
5	157+ mph	Catastrophic damage will occur with increased severity

Station detail is especially critical for ascertaining data completeness and quality, a realization erased by most summaries and grids. A common practice before 1982 was to assume missing daily quantity records were zero, a critical value for tracking drought. Thresholds cannot be reliably identified on days where values are left blank or assumed zero, unless, of course, a method is employed to generate values providing better information than the entered records.

With some R code already written and ready to run, delving into the data should be straightforward. The analysis is perhaps only as complicated as the weather.

2.2 The Source of Weather

A grasp of the concept of climate and its potential for change stems from understanding the physical source of weather: heat. The basics of daily and seasonal weather, which derive from heat and movement, explain the mechanisms of the El Niño Southern Oscillation, or "ENSO," with its varying phases of impact on regional climates.

The intense heat from the sun's rays near the equator seeks to equalize itself across the earth through winds and currents, all while the earth is engaged in two circular motions, rotating on a tilted axis while simultaneously orbiting the sun. People speak often of the "sunrise" and "sunset," and of changes in weather they feel which may be swift and drastic. Yet holding constant is the *imperceptive quality* of the underlying phenomena of motion around both an axis and a "stationary" sun, a sun that neither rises nor sets. While of little consequence to weather, the entire solar system including the sun and earth are actually moving through space around the Milky Way Galaxy in a third grander orbit. So the earth is in orbit, along with other planets, in a spiraling motion through space about a moving sun. The solar system's orbit might only impact the earth's climate over tens of millions of years. What is more, the Milky Way Galaxy is itself in orbit with other galaxies.

While essentially "sitting still" at her desk, an analyst could make fairly precise calculations of the earth's rotational motion based on latitude, all while feeling nothing of the earth beneath her speeding around and around at a staggering rate of over 800 miles per hour. This figure would increase to over 1,000 mph were she located near the equator. Simultaneously, she is orbiting the sun at 66,700 miles per hour so that in one full turn of the seasons, the distance traveled amounts to 584 million miles. The earth makes one complete revolution, completing a "sidereal day," in about four minutes short of 24 hours. Over four thousand miles of *orbit* completes each cycle of the "solar" day in an average of around four minutes – astounding speed! Since the earth's orbit it elliptical, the time and distance to complete a solar day varies with closeness to the sun. By the earth's dramatic motion in space, the state of heat inequality is driven by a rapid change of position.

With one half of the spheroid planet always illuminated, the surface of the earth travels *thousands* of miles in a single day's rotation to distribute heat evenly around it like a chicken roasting on a spit, which translates into a seemingly trivial differentiation of temperatures: cooler in the morning and at night compared to afternoon. The lag of several hours in respectively the warmest and coolest temperatures of the day following midday and midnight, comes from the magnificent ability of the earth's surface and atmosphere to store and slowly release heat energy. The *hundreds of millions* of miles in revolution through the solar system differentiates seasons – but only due to the slight tilt of the earth on its rotational axis. In its elliptical orbit, the earth's varying distance from the sun does not significantly influence temperatures. Rather the angle of the sun's rays decides intensity. A common illustration is a flashlight directed straight at a wall: moving the distance of the beam's source forward and away scarcely influences the light's intensity compared to angling its direction to a slant – the angle diffuses the brightness. Were the earth to spin straight up and down on a vertical axis while orbiting the sun, even hundreds of millions of miles could not produce a January distinguishable from June.

The atmospheric circulation on earth – a large scale movement of air distributing thermal energy across the earth's surface – can be described by the process of "convection." Convection is a circular motion of molecules within fluid, where fluids encompass both liquids and gases such as air. With a difference in temperature, hotter material rises while the colder sinks with gravity. In a room, hot air rises to the ceiling. On earth, the convective process occurs within the troposphere both latitudinally, from the equator to the poles, and longitudinally across the equator. From the equator to the poles, the decrease of solar intensity with latitude sets convective circulation patterns into motion. Along the equator, a difference in temperature arises between land and ocean because of the substantial difference in the amount of heat these surfaces types absorb and emit.

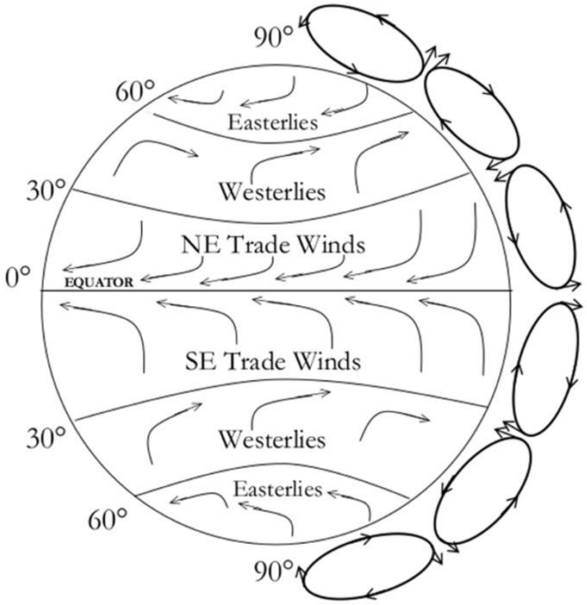


Figure 2. This illustration shows "idealized" patterns of ocean currents and the six convective cells which wrap around the globe within the troposphere, the lowest level of the atmosphere, where weather occurs. The rotation of the earth produces ocean currents flowing in opposite directions and breaks in the convective circulation loops at approximately 30° and 60° north and south.

Were the earth to stand still on its axis, cold winds would blow from the poles to the equator across its surface while hot air would rise at the equator in a convective circulation towards the poles. Rotation enters this equation with an elaborate influence, generating six segments of "idealized" wind directions that deviate from the theory, as all weather does, with changes in terrain and a profusion of random disturbances and interactions. Nearest the equator are the easterly (i.e. "from the east") trade winds which early merchant ships sailed, ranging from about 0° to 30° north and south. In both hemispheres from around 30° to 60° are the westerlies (i.e. "from the west") by which those ships made their return voyages, and at roughly 60° to 90° the circulation again reverses to easterly polar winds. Were it not for the complex circulation patterns arising from the earth's rotation, intercontinental trade could not have taken place by sail and oar. The force of the earth's rotation is strongest at the poles and weakens towards the equator, where seafarers could become trapped in the calm of the "doldrums." The circulation pattern along the equator, where rotation produces no force of deflection, is known as the "Walker circulation." At the equator, easterly winds across the wide open Pacific, in concert with the Walker circulation, give rise to the El Niño Southern Oscillation. Around the 30° latitude lines, subsiding dry air of the convective cells generates the desert regions in bands across Africa and Australia. From as far away as the farthest eastern end of the African deserts, dry subsiding air stirs winds that may continue to travel from east to west across the hot African land deriving strength to propel still further west across a warm Atlantic and morph into some of the most powerful hurricanes striking the eastern United States. This storm pathway illustrates the weather system is truly massive.

2.3 The Atmosphere

The atmosphere would be "paper thin" if the earth were scaled on the size of a basketball. The phenomenon of weather occurs only within its very base layer, the troposphere. Mount Everest, at just over 29,000 feet elevation (about five and a half miles), sits in the upper troposphere. The final layer of atmosphere ends about 6,200 miles from the surface which would only be a twelve hour flight, could an airplane traverse the thinning air.

Cold temperatures compress molecules, so that colder air is denser with less movement of molecules. Areas of high pressure – which essentially originate from coldness – move towards areas of low pressure – similarly defined by warmth – so the pressure differences from unequal heating near the earth's surface give rise to winds. The height of the troposphere varies with temperature and changes with seasons: at the equator it may extend as high as twelve miles while the winter poles may compress the layer to seven miles.

The earth's atmosphere is naturally comprised of gases. In dry air, without consideration of water vapor, the composition is roughly 78% nitrogen and 21% oxygen, gases which allow heat leaving the earth's surface to pass through and escape into space. The remaining roughly one percent of the mixture includes a very small proportion of "greenhouse gases" – gases typically measured in parts per million or billion which absorb heat released from the earth and trap them near the surface. These gases include carbon dioxide, methane and nitrous oxide. Water vapor is another greenhouse gas present in varying proportions by region, making up nearly 4% of the troposphere's gases in tropical regions near the equator, but closer to 1% near the poles. The proportion also varies through the natural cycles of cloud formation and precipitation. Without naturally occurring greenhouse gases, scientists estimate the average temperature at the earth's surface would drop from 59°F to 0°F. The mixture is precise: with less than 16% oxygen content, ordinary fires would not burn; while high oxygen concentrations would aggravate combustibility. Therefore these molecular elements are precious to life on earth, and no more detectable to us than the motion of the earth beneath our feet. Yet imagine in its entirety, only a few miles outside the range of sight and rotating along with us, this thin invisible atmosphere is enough to disguise the hurling high speeds of the earth's rotation and orbit! This illustrates that the climate system, while massive, is also meticulously detailed.

Scientists agree that adding greenhouse gases to the atmosphere will raise surface temperatures. The warming effect of recent history is best illustrated in the award-winning documentary "Chasing Ice" in which photographer James Balof chronicles the rapid melting of glaciers. Charles Keeling began recording carbon dioxide levels in the atmosphere at Mauna Loa Observatory beginning in 1958, noting seasonal variations of concentrations in the atmosphere; by 1961 he issued the first warnings of anthropogenic contributions to the greenhouse effect. Roger Pielke Sr. stirred controversy in 2007 by claiming carbon dioxide accounts for only 28% of human-caused warming, stressing the remaining 72% is still human caused.

Large bodies of water absorb and release heat at a much slower rate than the atmosphere or ground terrain, requiring over a thousand times the energy to heat as the same volume of air. The upper ocean near the surface can store approximately 30 times the heat as the atmosphere immediately above it. Interaction between water bodies and the atmosphere also creates sea breezes. These phenomena lead climates near coasts and large lakes to be more temperate than areas inland. The ocean is a gigantic sink for atmospheric warming, the effects of which may not be felt so well on land until the ocean has reached its full capacity for absorption.

Other human activities and natural forces can cause temperatures to rise, or fall, and climate change collectively refers to all types of changes to regional climates or long-term weather patterns and extremes, not only heating, but cooling or changes in precipitation or winds. For instance, deforestation releases carbon to the atmosphere but further alters surface reflectivity from greener to drier while removing the valuable balancing process of photosynthesis by which carbon dioxide is converted with sunlight into oxygen. Forests can suddenly be replaced by agriculture or housing tracts; water use, land use, and controlled burning can all immediately influence climate. City streets of asphalt have induced the "urban heat island effect," an effect that can be counteracted with the numerous benefits of roof gardens. Nuclear power plants raise the water temperature of adjacent lakes that supply water to cooling towers. Natural volcanic eruptions spew carbon and particulate matter into the atmosphere, typically cooling the earth for several years from the high reflectivity of particles. Particulate matter from all types of pollution, even dust rising from cleared fields, assists storm clouds to grow larger and form into more powerful storms. While greenhouse effects are described as slow and gradual, many types of climate change are more immediate including the natural cycles of ENSO.

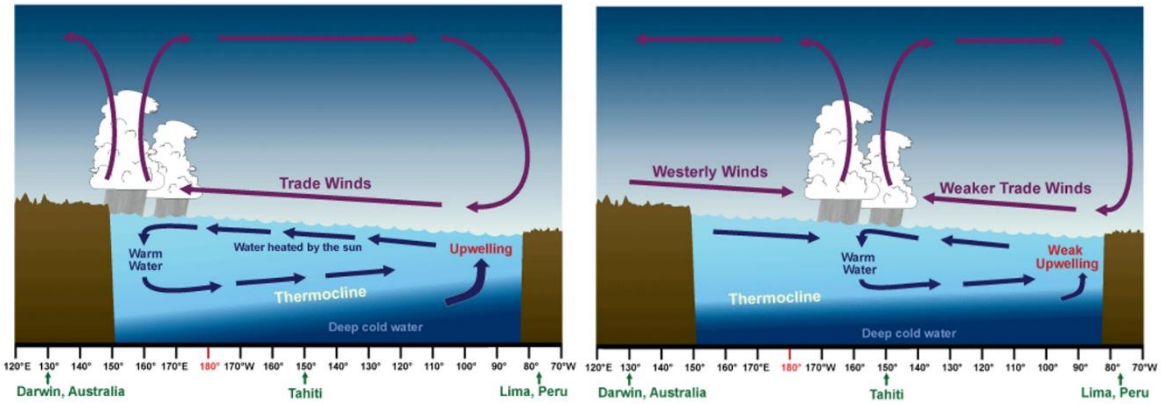
2.4 El Niño Southern Oscillation

The Pacific is the largest body of water in the world, twice as large as the Atlantic and far deeper. Its expanse across the hot equatorial region wraps nearly half the earth's circumference, spreading the canvas for the brush strokes of the El Niño Southern Oscillation, or simply ENSO. Temperature and pressure will typically differ substantially from one end of the Pacific to the other. The tropics of the western Pacific hold some of the hottest water in the world's oceans: surface temperatures may warm to around 84°F covering an area the size of Australia. At the Peruvian coast, temperatures may be as cold as 60°, uncharacteristically low for the tropics. Yet the sun's rays are of equal strength all across this equatorial region.

Motions and attributes of oceans are not separated from atmosphere; rather the two interact with "positive feedback loops" by which changes are amplified, pushing away from equilibrium to invite instability. The atmosphere responds to disruptions quickly in time scales of days to weeks, while the ocean reacts more slowly, over months to years. The El Niño Southern Oscillation is a single large-scale coupled interaction of atmospheric pressure and ocean temperature across the Pacific Ocean, stretching from the coast of South America at Ecuador and Peru in the east to Indonesia and Australia in the west. "Southern Oscillation" refers to the "seesaw" effect in atmospheric pressure between the eastern and western Pacific: when pressure at one end shifts to lower than normal the other end will become higher than normal. "El Niño" refers to ocean warming across the Pacific equator which occurs together with the dominating shifts of pressure. These shifts in the tropics can exert powerful influence on global weather.

Beneath an evenly intense sun, a striking contrast in surface temperatures arises at opposite ends of the central Pacific. Five major contributors emanate primarily from the earth's rotation: (1) heating by both sun and warm air as water is pushed westward along the equator by trade winds, (2) upwelling along the equator by the same motion of the trade winds, (3) cold upwelling at the coast of Peru, (4) warm downwelling at Indonesia, and (5) the change in the depth at which colder waters lie from the surface, across the equator. The underlying mechanisms deserve elaboration before considering how a reversal takes place.

As winds blow across the surface of any body of water, the turning motion of the rotating earth will cause the water to spiral so that it moves overall perpendicularly to the wind direction. In the southern hemisphere, water is deflected to the left of the wind direction; and in the northern to the right. Winds blowing towards the equator from both the north and the south turn towards the west. As water is displaced along the equator from either side, water from below the surface rushes in to replenish the space. Winds blowing northward along the coast of Peru similarly produce an upwelling, where temperatures near the ocean's surface are cold. The opposite occurs at Indonesia and the other land barriers of the Maritime Continent, where westward winds produce a downwelling of warm surface waters according to the direction of the earth's rotation. Sinking warm waters push the colder basin waters below down to even further depths. Note that the ocean is stratified: water near the surface is warm from various influences such as the sun's heat, evaporation, and mixing winds; while deeper waters are still and cold. The "thermocline" lies in between, a thin dividing layer in which temperatures drop quickly through a shallow depth. The sinking of warm waters at the western Pacific encourage a downward slope to the thermocline from east to west. The waters upwelled along the equator by "the trades" increase in warmth moving west as the cold lower layer slopes down further and further below surface.



Source: NOAA Jetstream

Figure 3. El Niño Southern Oscillation. The 'normal' state of the Pacific Ocean is illustrated on the left; but when conditions are amplified the same pattern become a La Niña event. An El Niño event is illustrated on the right.

Warm surface water pushed westward by the trades eventually encounters barriers in the land masses of Australia and Indonesia, where it literally piles up. Over time the western sea level may gain 20 inches elevation, forming a mound of water visible from space. This view serves as an assessment of the ENSO phase. The slope of the ocean's surface, then, opposes the slope of the thermocline. The heated water that reaches the Maritime Continent of Indonesia evaporates from the ocean, condenses into rain clouds, and pours out tropical rain storms, fueling upper level winds. Every year, over 100 quadrillion (10^{17}) gallons of water evaporates from the ocean, mostly around the tropical equator, with about 90% of the precipitation falling over the ocean. Rising warm air travels through the troposphere eastward back across the equator and then settles in a convective loop, reinforcing the westward trade winds along the surface.

Awareness of a reversal in the usual pattern originated in Peru. Ordinarily, winds blow northward along the coast of Peru stirring up cold waters, replacing depleted surface waters by rich nutrients from deep basin waters – that feed vibrant fish populations – which in turn sustain bird populations – whose droppings provide fertilizer to the agricultural sector. A seasonal transformation of an inconsistently warm current entering this coastal region was first identified by Peruvians at Christmastime as El Niño, *the boy* or *the Christ child*. A strong El Niño event can devastate Peruvian fisheries, impair agriculture, and induce rain storms that flood the coastal regions.

When trade winds are brisk, coastal upwelling is strong along Peru, and the thermocline is steep, an amplified phase of colder eastern sea surface temperatures may be referred to as La Niña, or *the girl*. The same conditions at a lesser strength are considered "neutral," or the "normal" state of the Pacific – sometimes called La Nada, *the nothing* – a state which does not prompt severe weather.

An El Niño event always begins with pressure changes, namely, a lessening of the pressure gradient between the eastern and western Pacific. Since winds blow from high to low pressure, this leveling of pressure weakens the trade winds that have driven water to pile up towards the west. The heated water will then slosh back in a countercurrent that sends the excesses of warm water across the Pacific. The central and eastern regions of the Pacific waters warm near equal to the western temperatures, repositions the intense rainstorms away from Indonesia towards the central or eastern Pacific, and shifting large scale wind patterns in turn. Pronounced phases of ENSO – El Niño and La Niña alike – are known for diverse consequences of extreme weather at near and distant regions of the globe, *sometimes* with opposite impacts to one another. All of the effects together do not amount to true opposites considering some arise from a shift in the region of predominant precipitation, a location change which is not an opposite.

The term "El Niño" has come to signify an amplified cycle which typically occurs on intervals of three to five years, historically from two to seven years. Variation is not only in frequency and strength but also duration which may span several months to a few years. La Niña is especially well known for enhancing Atlantic basin hurricane activity. Within the troposphere where weather occurs, various wind speeds and directions may occur for several miles above the ground known as "vertical wind shear" which when strong, can topple hurricanes or stifle their formation. La Niña conditions foster an evenness along altitudes favorable to hurricane formation and survival, that is, a weakening of vertical wind shear. Other consequences of El Niño and La Niña are shown in the maps following.

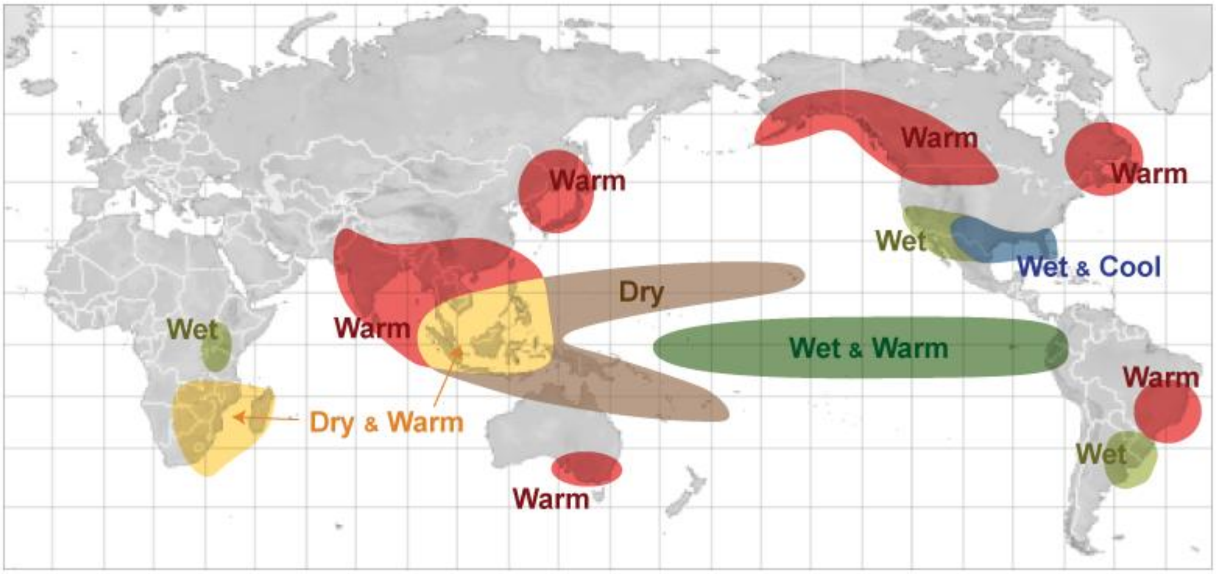
What triggers the pressure gradient to lessen, unleashing an El Niño event, remains a scientific mystery. There may not be one precise answer since weather is influenced by numerous factors characterized by random occurrences, and further by interactions and also feedbacks. Ambient air pressure is constantly changing, and even while the pressure changes are measurable, the sources of change may not be discernible. Random distortions to any number of usual weather patterns or combinations thereof could eventually lead to shifts of pressure at the equator: sudden bursts of opposing winds, sub-surface waves, changes in salinity from the sinking of salty waters along the equator, or distant elements such as mountain snowpack or glacier ice, could shape valid hypotheses. This mystery beneath recurring large-scale global weather patterns illustrates that the climate system, both massive and detailed, remains largely "over our heads."

Table 2. Summary of ENSO event characteristics – the phases of the El Niño Southern Oscillation may be summarized by a few characteristics. When La Niña conditions are present but are mild, not amplified, the phase is neutral and global weather patterns are not influenced.

La Niña	El Niño
Strong upwelling of cold deep basin waters at coast of Peru	Weaker upwelling along Peruvian coast, and upwelling of warmer waters
Steep thermocline with cold water nearer to surface	Less slanted thermocline
Strong easterly trade winds	Weakening easterly trade winds
Warm western Pacific and cooler eastern and central Pacific	Central to eastern Pacific assume warmer temperature, nearer that of western Pacific
Region of persistent precipitation is over warmest water near Indonesia	Region of persistent precipitation is shifted, over warmest water near central Pacific
High sea level pressure in eastern Pacific differs from low pressure in western Pacific – strong Walker circulation	Sea level pressure in eastern Pacific lowers near to level of western Pacific – weakening Walker circulation

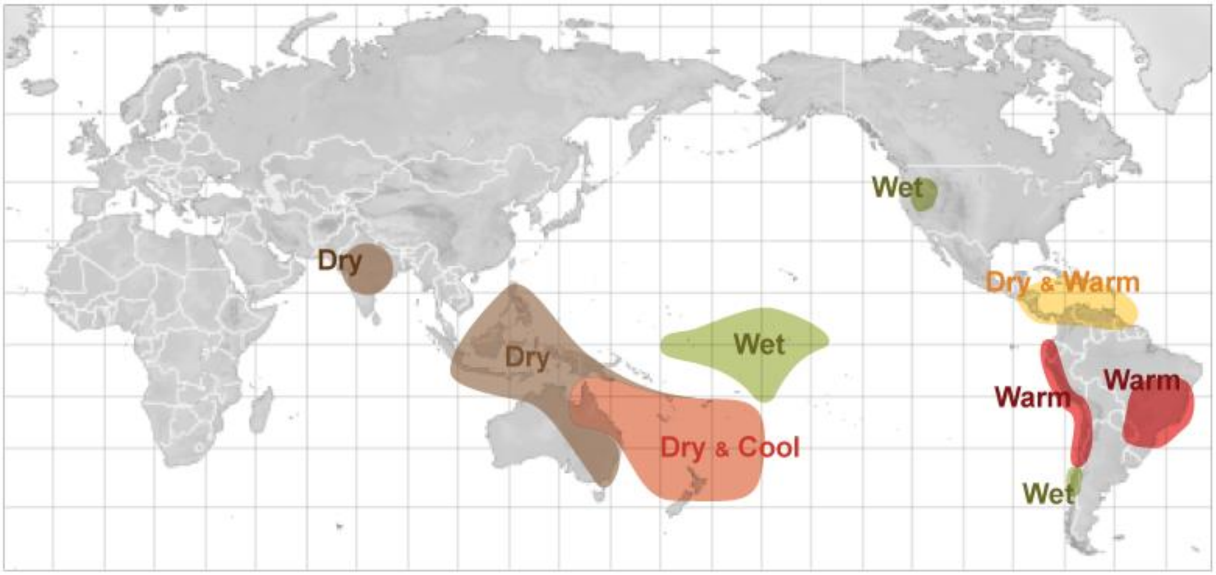
Figure 4. Regional Weather Impacts of El Niño Southern Oscillation – El Niño and La Niña – winter and summer seasons. These four maps provided by the NOAA serve as excellent reference to the regional effects of natural climate cycles in weather data specific to the ENSO phenomenon, and may be consulted for planning climate phase analyses by location and time of year.

(A) El Niño - winter season
El Niño effects during December through February



Source: NOAA Jetstream

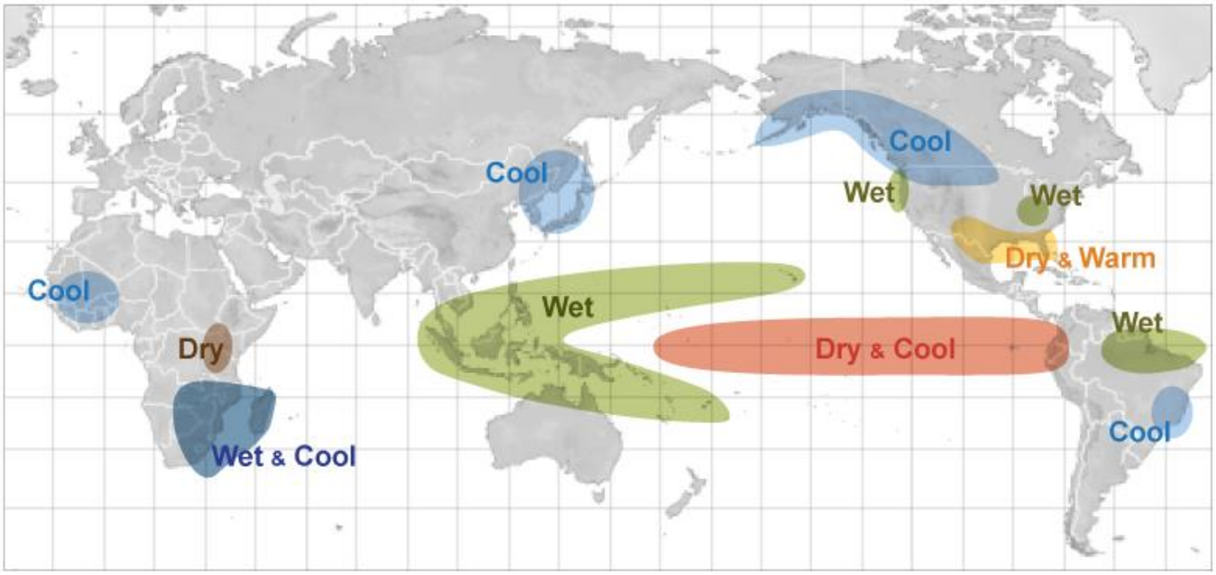
(B) El Niño - summer season
El Niño effects during June through August



Source: NOAA Jetstream

(C) La Niña - winter season

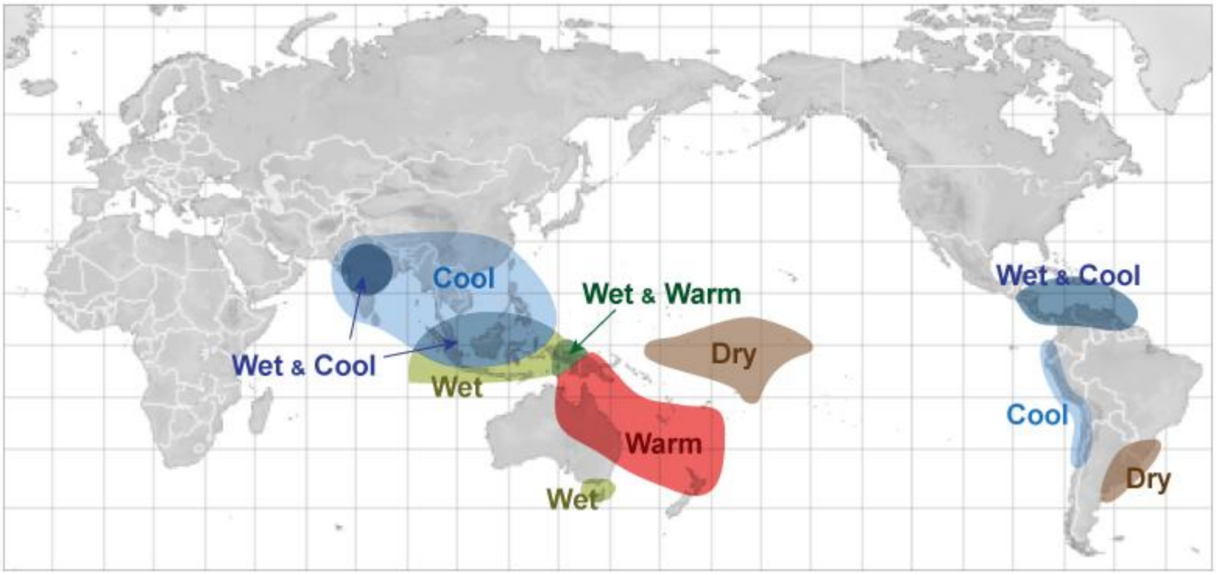
La Niña effects during December through February



Source: NOAA Jetstream

(D) La Niña - summer season

La Niña effects during June through August



Source: NOAA Jetstream

2.5 ENSO Indices

The original indices tracking the phase of ENSO are named by the ship tracks that originally recorded sea surface temperatures (SST) across this equatorial region of the Pacific, beginning with Niño 1 and 2 near the coast of Peru where destructive forces of the ENSO phenomenon were first witnessed. Niño 3 extends across an eastern equatorial band of the Pacific, reflecting the later realization of a farther reaching phenomenon. Niño 4 covers the tropics to the west, and Niño 3.4 is measured in a midregion overlapping Niño 3 and 4. The Niño indices are recorded most commonly as average monthly SST and are also given weekly, and further as anomalies from a base mean SST value. The more extreme colder temperatures relate to La Niña events, the warmer to El Niño.

The Ocean Niño Index (ONI) is derived from the Niño 3.4 SST as rolling three month periods (Jan-Feb-Mar, Feb-Mar-Apr, etc.). The Trans-Niño Index (TNI) is derived in a different manner combining Niño 1 and 2 with Niño 4. The TNI considers that the difference in SST on opposite sides of the Pacific better reflects the phase for certain purposes, and takes the standardized Niño 1 and 2 minus the Niño 4 with an additional standardizing adjustment; specifically, a five month running mean is applied and then standardized using the 1950-1979 period. The regions of measurements for ENSO indices are shown in the map below.

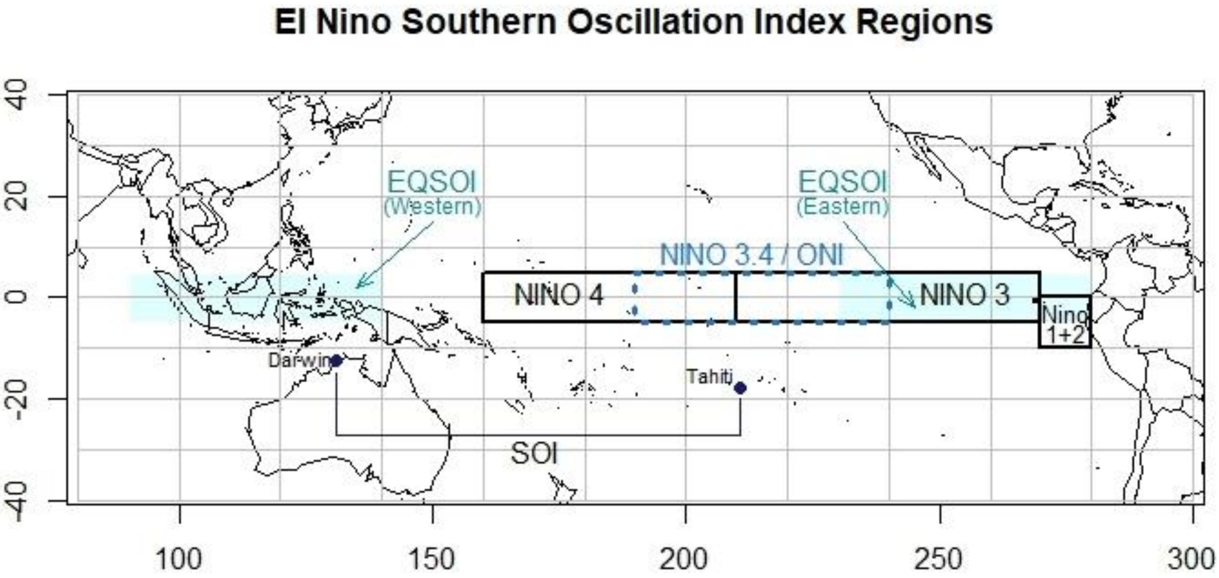


Figure 5. ENSO Regions – regions where the phase of ENSO is measured by SST or pressure are shown in a Pacific-centric map. The TNI is based on Niño 1+2 and Niño 4 while 'BEST' is based on Niño 3.4 and the SOI.

Table 3. ENSO Index Coordinates – the coordinates where ENSO indices are measured are given in Atlantic-centric coordinates (-180° to 180°) and Pacific-centric coordinates (0° to 360°)

ENSO Index	Atlantic Coordinates		Pacific Coordinates	
Niño 1+2 / TNI (east)	0°-10°S,	90°W-80°W	0°-10°S,	270°E-280°E
Niño 3	5°N-5°S,	150°W-90°W	5°N-5°S,	210°E-270°E
Niño 3.4 / ONI / 'BEST'(i)	5°N-5°S,	170°W-120°W	5°N-5°S,	190°E-240°E
Niño 4 / TNI (west)	5°N-5°S,	160°E-150°W	5°N-5°S,	160°E-210°E
EQSOI (west)	5°N-5°S,	220°W-270°W	5°N-5°S,	90°E-140°E
EQSOI (east)	5°N-5°S,	80°W - 130°W	5°N-5°S,	230°E-280°E
SOI / 'BEST'(ii):				
Darwin, Australia			12.4634°S,	130.8456°E
Tahiti			17.6509°S,	210.5740°E

Further indices exist to track the ENSO phase without SST measures. The Southern Oscillation Index (SOI) records the large-scale fluctuations in pressure between the western and eastern Pacific, at the locations of Darwin, Australia versus Tahiti. The pressure differential is associated with heat in the atmosphere as opposed to the surface water of the ocean, and the atmospheric pressure gradient is prone to change much more swiftly than ocean temperatures. The SOI is more negative during an El Niño event, where pressure in the eastern Pacific lowers nearer to that of the western Pacific. The Equatorial SOI is another measure based on pressure, but instead of relying on two distinct points observes averages across larger regions, over Indonesia and off the coast of Ecuador.

The Multivariate ENSO index (MEI) combines several characteristics into one index. Its calculation considers the six main observed variables over the tropical Pacific: sea-level pressure (P), zonal (U) and meridional (V) components of the surface wind, sea surface temperature (S), surface air temperature (A), and total cloudiness fraction of the sky (C); calculated in rolling bimonthly periods (Jan-Feb, Feb-Mar, etc.). Various index measures track different characteristics of the ENSO phase, so they will serve as unequal indicators to climate effects in various regions of the globe. Klaus Wolter of the NOAA describes the relevance of the MEI, in relation to other indices, as follows:

> "Why do I believe that the MEI is better for monitoring ENSO than the SOI or various SST indices? In brief, the MEI integrates more information than other indices, it reflects the nature of the coupled ocean-atmosphere system better than either component, and it is less vulnerable to occasional data glitches in the monthly update cycles. Now, if you are interested in ENSO impacts in a very specific part of the world, I would suggest that you obtain other ENSO indices as well and establish which one best fits your needs. For instance, in Australia, Darwin sea level pressure and/or the SOI may be more appropriate than the MEI. My claim here is that the MEI does a better job than other indices for the overall monitoring of the ENSO phenomenon, including, for instance, world-wide correlations with surface temperatures and rainfall."

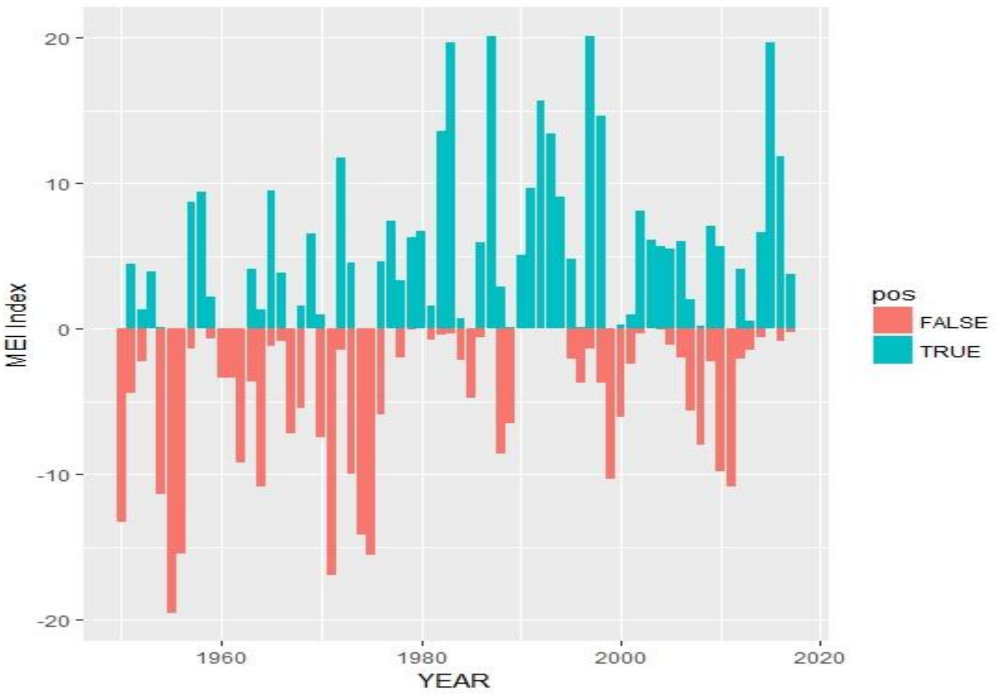


Figure 6. Phases of the MEI. Multivariate ENSO Index since 1950.

Indices tracking ENSO phases are available online at these NOAA sites:

	Website Address	Indices (format: Wide or Long)	from
(I)	www.cpc.ncep.noaa.gov/data/indices/	Niño, ONI (L); SOI, EQSOI (W)	1950's
		Niño Weekly (L)	1990's
(II)	www.esrl.noaa.gov/psd/gcos_wgsp/Timeseries/	Niño (W); SOI (W)	1870's
(III)	www.esrl.noaa.gov/psd/data/climateindices/list/	Niño, ONI SOI, TNI, BEST, MEI (W)	1950's
(IV)	www.esrl.noaa.gov/psd/enso/mei/table.html	MEI (W)	1950's

(I) NOAA National Weather Service – Climate Prediction Center – Monthly Atmospheric & SST Indices
(II) Global Climate Observing System – Working Group on Surface Pressure

NOAA– Earth System Research Library – Physical Sciences Division – Climate Indices – Monthly Atmospheric and Ocean Time Series
(III) NOAA– Earth System Research Library – Physical Sciences Division – Multivariate ENSO Index

2.6 Climate versus Weather

Weather typically describes short-term phenomena while climate describes the long-term weather conditions that predominate a specific region. A "climatological normal" is an average of a weather element over 30 years, which serves as a base for comparison. For scientific purposes, climate is usually defined by a 30-year period; for some purposes, the base climate period chosen might span 40 to 100 years. The definition of climate includes not only the long-term averages and typical variations in the elements, but further places emphasis on the extremes experienced over the full range of the selected base period. A "very hot day," then, describes weather, while "the hottest day in London since 1976" designates a boundary for one major city's climate.

The common 30-year scope implies that weather is expected to fluctuate to a certain extent, from one year to the next, and variations in this range would not constitute climate change. The assumption that three decades would cover the irregular fluctuations of the El Niño Southern Oscillation might also be implied, since this is the major cyclical climate factor for some regions. But because of the myriad of interactions among climate variables, not only is ENSO a source of natural climate variations, ENSO is itself susceptible to change. A base climate period might be more closely examined for trends, cycles, and shocks. Irregularities might be taken to another level of comparison and adjustment when considering future loss potential.

For the examples of this paper, the years 1961-1990 are selected as a base period for climate. This period corresponds to the earliest 30-year term at which instruments are considered reliable and consistently gauged. Care should certainly be taken in relying upon analyses which include decades prior to the 1960's since old ship records or primitive instruments may reflect not a change in climate but rather a change in measurement capabilities or variations in techniques for capturing data.

Certain adjustments to daily data will remain essential since the 1960's, due to inconsistencies in recording zero measurements, or the closing and opening of weather stations, for instance. Changes in data quality have been especially drastic since 1982 as a range of improvements were implemented for achieving more complete, more consistent records. Some of the prominent data changes are presented in summary in the 'Results and Discussion' section.

2.7 Actuarial Analysis

Weather is no stranger to the insurance industry; policies insuring ships against storms and other causes of sinkage were first written Before Christ. Modeling weather has become a standard only since Hurricane Andrew, and still catastrophe simulations are proprietary which limits discussion beyond what little the model designers and their clients wish to share. The duration of property insurance policies rarely exceeds one year, so insurers can adjust premiums in response to gradual, long term climate mechanisms and may not need to discern source changes. Primary consideration might be given to ENSO phases, which can be predicted sometimes six months in advance. Other short-horizon climate disruptors may possibly receive some attention. Yet, with growing concern over nature's destructive forces, the role of weather and risk experts may need to be updated to include more than the offering of near-term insurance policies.

Actuaries possess refined comprehension of the messages raveled inside vast sets of data. The need to measure economic costs of calamities has given actuaries a uniquely precise viewpoint of risk assessment. Actuarial science can bring advancements to climate analysis in such areas as credibility and outliers, treatment of sparse data, recognition of interactions, removal of double-counting, identification of noise signals, normalization, trending and pattern searching. Actuaries have placed greater focus on mathematical aspects of storm losses and are far more rigorous in these numerical areas than the other sciences. The treatment of catastrophic weather loss in models combines the skills of the actuary with the atmospheric scientist, together but separately, in a limited market. Techniques in weather and catastrophe may be progress to apply financial and actuarial expertise directly, along with the distinct qualities of physical sciences.

3. RESULTS AND DISCUSSION

Results are given from output of the code provided in the appendix, and serve as examples of the much wider range of information the meteorological data sets can provide.

3.1 Data Completeness

Stations open and close over time, with changes of location; differences in elevation and surroundings impact measurements. While precipitation (PRCP) has been recorded at over 56,000 stations in the United States and Canada since 1960, fewer than six percent of these stations contain data for 30 base years and the subsequent 27 years for comparison.

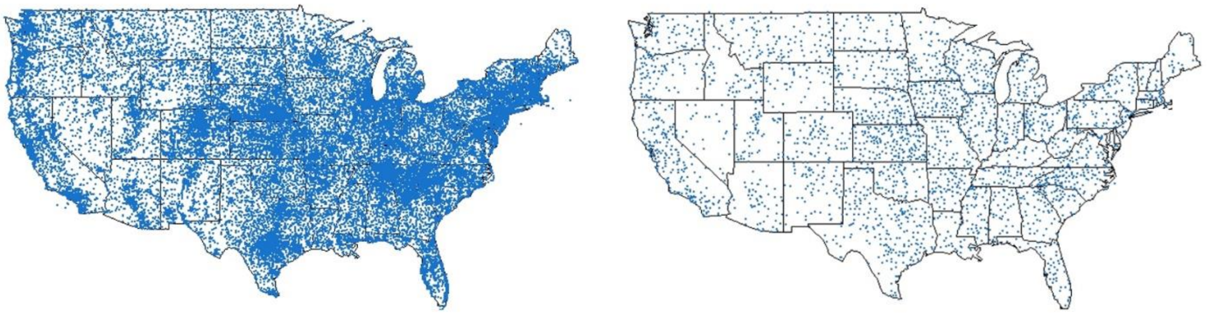


Figure 7. Precipitation (rainfall) records have been recorded at over 48,000 stations in the United States since 1960 (left figure). Only 6.4% of these stations records include some data in all 57 years from 1961 to 2017 (right figure); however, over 26% of the yearly precipitation data for these decades was recorded at these long-operating stations.

The GHCN-D data is fraught with missing records. Beginning in 1982, an existing notation became commonly utilized to indicate a blank that had been assumed zero, for quantity measures such as rainfall. The number of identifiable missing records jumped in 1982, and new initiatives were taken so that record completion has improved since then. The practice of assuming zero records was phased out by the end of 2010. Prior to 1982, blanks that were assumed zero cannot be identified, so while the data appears more complete for older years, in reality, the zero records are unreliable.

The National Centers for Environmental Information (NCEI) of the NOAA also provides monthly GHCN data summaries of weather elements (GHCN-M), to which 'homogeneity adjustments' have been made [www.ncdc.noaa.gov/ghcnm]. The online data source includes reference materials describing adjustments that are called for by the raw daily data records.

Table 4. Change in assumed zeros. In 1982, GHCN-D missing records appeared to increase only because blanks became identifiable by notation; subsequently completeness has improved.

Missing Records as a % of days of year

Year	PRCP	SNOW	SNWD	TMAX	TMIN
1960	2.8%	4.0%	6.2%	2.6%	2.7%
1961	3.2%	5.4%	8.5%	3.2%	3.3%
1962	3.2%	5.5%	8.0%	3.3%	3.3%
…					
1979	5.7%	8.5%	11.2%	7.0%	6.8%
1980	5.4%	8.8%	11.5%	6.8%	6.8%
1981	4.1%	6.9%	8.6%	5.8%	5.8%
1982	31.4%	82.6%	78.6%	4.7%	4.5%
1983	31.3%	83.0%	78.9%	4.2%	4.1%
1984	32.0%	83.5%	77.7%	4.6%	4.6%
…					
2015	21.1%	41.7%	44.6%	4.5%	4.6%
2016	22.0%	39.5%	41.3%	4.7%	4.8%
2017	19.4%	34.4%	32.7%	4.4%	4.4%

Zero observations* as a % of observations

Year	PRCP	SNOW	SNWD	TMAX	TMIN
1960	75%	96%	90%	9%	34%
1961	73%	97%	92%	7%	33%
1962	74%	97%	91%	8%	33%
…					
1979	73%	96%	90%	10%	35%
1980	75%	97%	92%	8%	34%
1981	74%	97%	94%	7%	33%
1982	60%	78%	58%	9%	34%
1983	60%	79%	62%	9%	33%
1984	61%	80%	63%	8%	34%
…					
2015	68%	96%	79%	8%	31%
2016	69%	96%	82%	8%	31%
2017	68%	96%	82%	8%	32%

* For temperatures, 'zero observations' are counts at or below zero degrees Celsius (freezing temperatures).

It might be expected that rainfall (snowfall) might not be recorded reliably during extremely dry (hot) weather. For snowfall in Minnesota, missing records average 45% for summer months for which all records are zeros, but still over 25% of records are missing in snowy winter months.

Table 5. United States Precipitation Records. The percentage of daily records each year seems to be falling while actually data quality is improving. Prior to 1982, blank records were assumed zero but most lacked identifying notation.

			% days				
	% days	% days	blank			blank	
	with	record	assumed	% days	% records	assumed	count of
Year	record	missing	zero	zero	zero	zero	stations
1961	95.1%	3.2%	0.1%	69.7%	73.3%	2,343	9,704
1962	92.2%	3.2%	0.1%	68.7%	74.5%	2,291	9,757
1963	94.6%	3.3%	0.1%	72.3%	76.4%	2,762	9,479
...							
1979	92.7%	5.7%	0.1%	67.6%	72.9%	2,447	8,527
1980	92.1%	5.4%	0.1%	69.2%	75.1%	2,477	8,650
1981	94.2%	4.1%	0.1%	69.6%	73.9%	2,658	8,690
1982	67.3%	31.4%	25.8%	40.2%	59.7%	817,610	8,673
1983	67.7%	31.3%	26.4%	40.4%	59.7%	833,783	8,646
1984	67.0%	32.0%	27.0%	40.9%	61.0%	851,171	8,608
...							
2008	69.3%	20.0%	5.8%	47.5%	68.5%	428,242	20,058
2009	70.6%	20.1%	3.9%	47.3%	67.0%	323,319	22,537
2010	71.9%	21.1%	3.1%	50.1%	69.7%	268,530	23,497
2011	71.8%	21.1%	0.0%	50.0%	69.6%		24,411
2012	72.2%	20.4%	0.0%	51.9%	72.0%		25,516
2013	71.9%	20.9%	0.0%	49.8%	69.3%		26,427
2014	71.7%	21.4%	0.0%	49.3%	68.9%		26,366
2015	72.7%	21.1%	0.0%	49.2%	67.6%		26,017
2016	73.9%	22.0%	0.0%	51.1%	69.2%		24,381
2017	72.5%	19.4%	0.0%	49.5%	68.3%		25,605

The change in missing records is explained Dr. Matt Menne, the creator of the GHCN-Daily meteorological databank at the NOAA's National Centers for Environmental Information (NCEI):

> "Many volunteer observers, especially in the more historic past, have not consistently recorded zeros each day when no rain was observed and rather would often leave the day blank in such cases. Because zeros have so often been left blank on reporting forms, NCEI used to more or less routinely assign a zero value to daily precipitation totals that were left blank. These added zeros were intended to be accompanied by a flag noting that the value "was missing but presumed zero" so that they could be distinguished from days when the observer noted a zero. However, the practice of assuming zeros for blanks was discontinued after 2010 when we moved to a new ingest and processing system for daily data, largely because the accuracy of assuming a zero for blanks could not be assessed very well. In addition, volunteer observers were rapidly transitioning to electronic reporting around the same time and are now prompted somewhat by the new electric entry system as to whether a missing value was really meant to be reported as a zero."

3.2 Choropleth Maps

Choropleth maps are color coded ranges that allow immediate visual interpretation. R contains numerous packages that will produce a choropleth map, although most are designed for quantity measures and lack flexibility for other purposes. The example choropleths plot anomalies centered at zero, which is straightforward to code in the package 'ggplot2' but may be more cumbersome to produce with other packages. The package 'ggplot2' has an advantage of being compatible with 'fiftystater' that includes insets of Alaska and Hawaii.

The first choropleth example is created from scratch in package 'maps' and provides code that allows for a high degree of customization. A drawback of this package is the lack of insets for Alaska and Hawaii, although these states can still be mapped separately.

The code allows for a year to be selected, which is compared against the base climate period (1961-1990). The base period average and standard deviation are calculated for each state separately. The choropleth shows the number of deviations upward or downward from the base average.

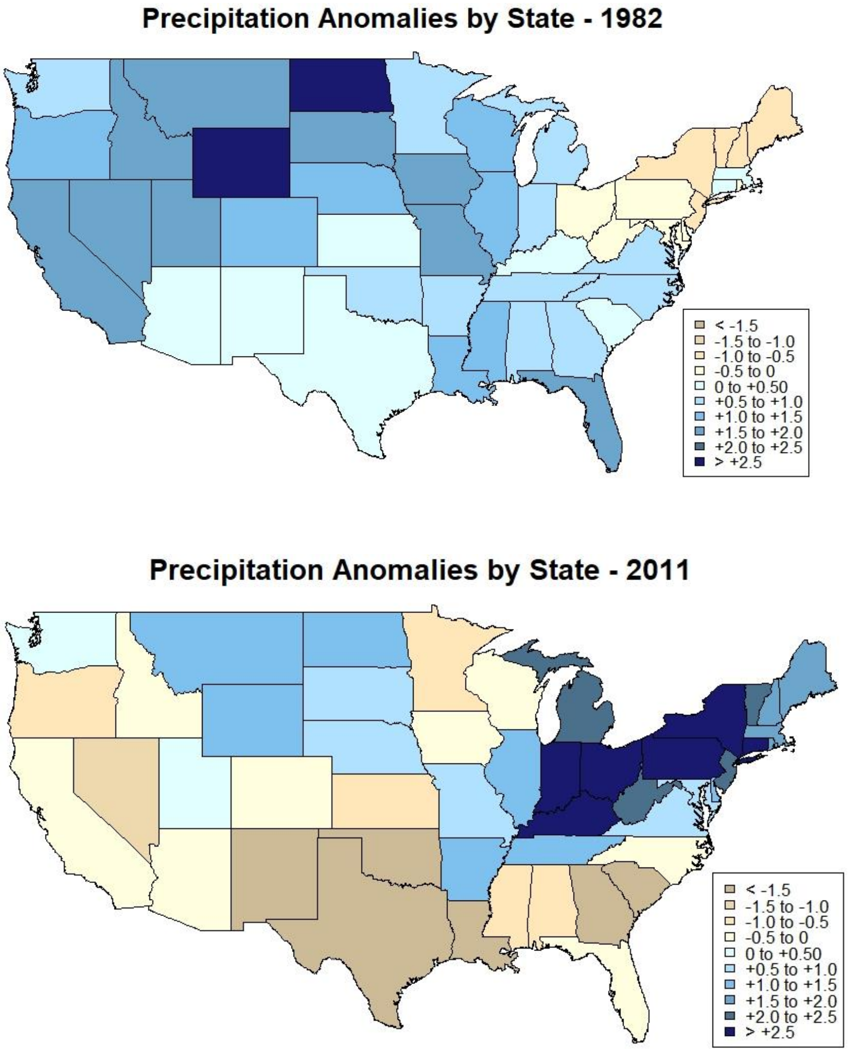


Figure 8. Choropleth maps produced from scratch using package 'maps' for an El Niño event year 1982 (top) in contrast to a La Niña event year 2011 (bottom).

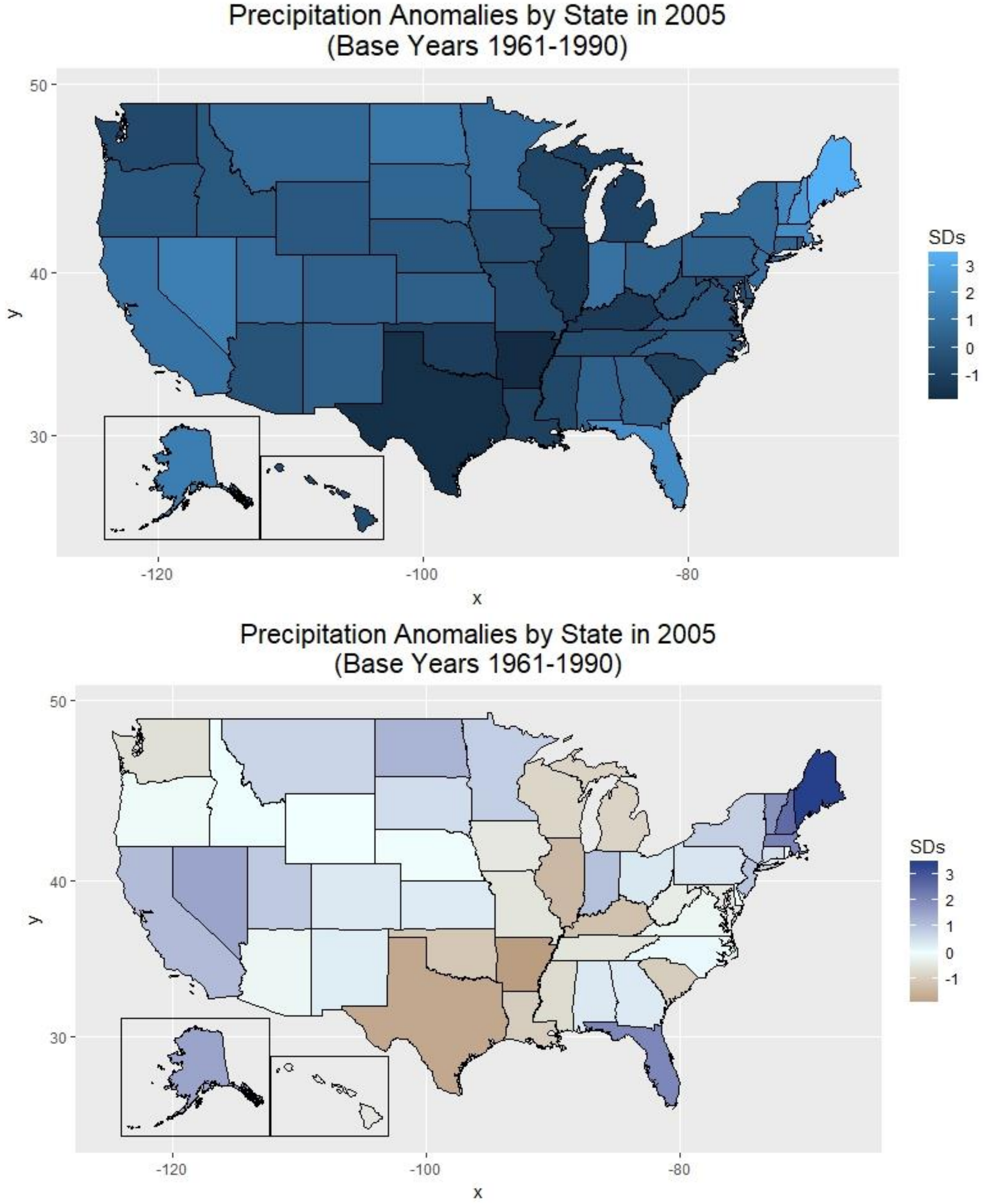


Figure 9. Choropleth map produced by package 'ggplot2' with package 'fiftystater' insets, using the base color scheme (top); and using a custom color scheme with a midpoint specified at zero anomaly (bottom).

3.3 Plots of Elements vs. Indices

Florida winter precipitation (PRCP) is chosen as an example region from the NOAA Jetstream maps, which indicate wet and cool conditions are expected during El Niño phases, dry and warm during La Niña. Several ENSO indices and time periods are selected to plot against the average daily recorded rainfall. Only stations have been included with some data in all 57 years (1961-2017); data completeness by month has not been checked. No adjustment has been made for assumed zero entries prior to 1982 which lack notation as blank records. The plots assign a shape to distinguish points in the two decades before 1982 which could adjust upwards due to an over prevalence of zeros.

For the Niño indices, there does not appear to be a strong relationship. For the MEI, the correlation with Florida rainfall appears convincing from January to March, but not in December. By this example, the choice of index would appear critical for identifying the specific characteristics of ENSO that impacts the region. If a loss threshold has been established for Florida rainfall, then a relationship between the MEI and the threshold might cause an insurer to consider ENSO phases in its loss history and realign expectations for the future.

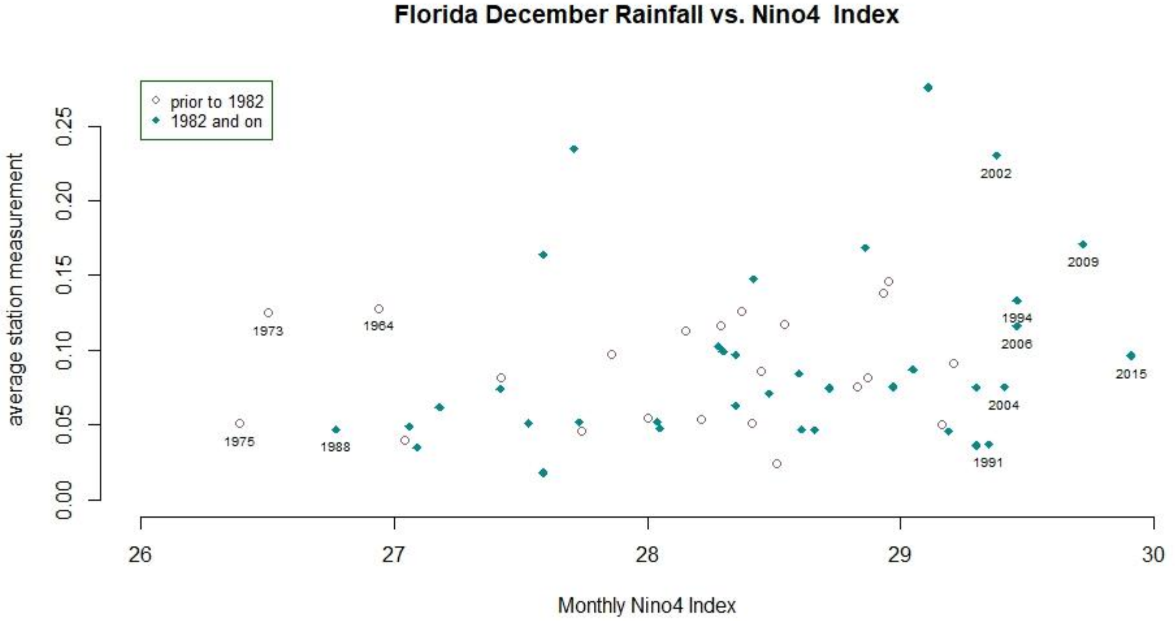


Figure 10. Florida rainfall (PRCP) in December plotted against the Niño 4 Index does not reveal a distinct pattern.

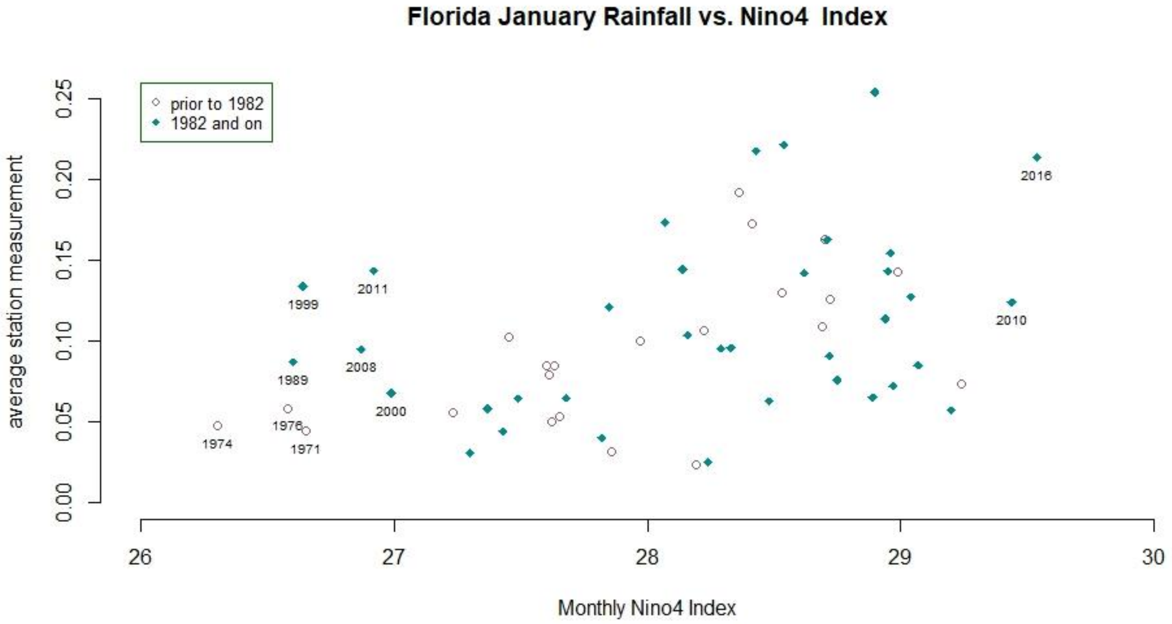


Figure 11. Florida rainfall (PRCP) in January plotted against the Niño 4 Index does not reveal a strong phase relationship.

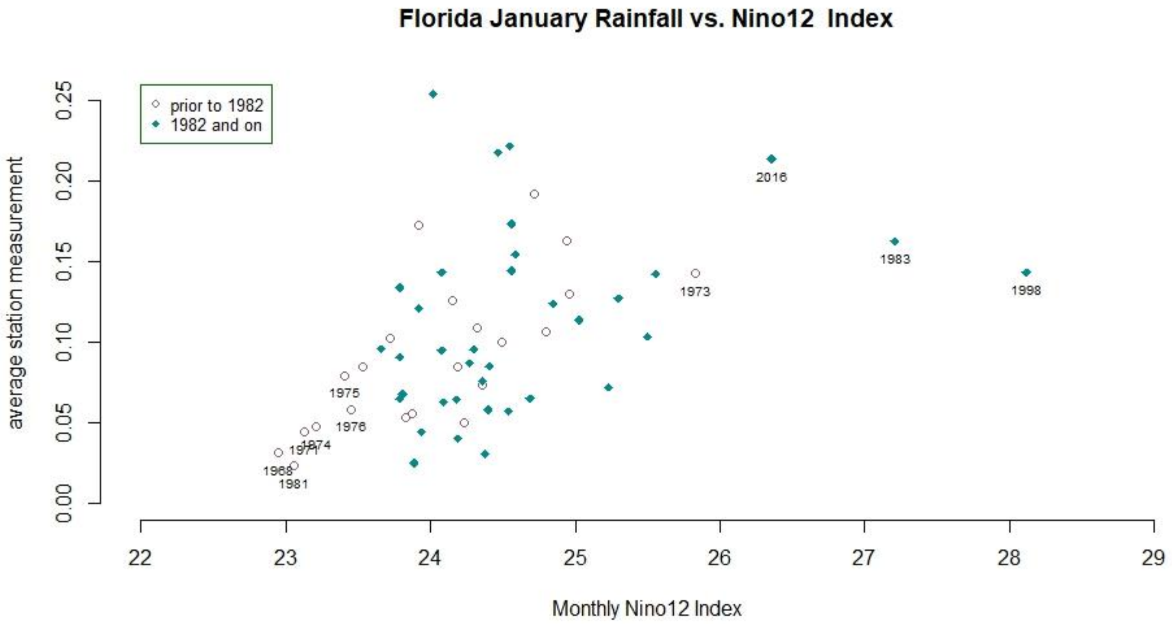


Figure 12. Florida rainfall (PRCP) in January plotted against the Niño 1+2 Index might reveal a slight phase relationship.

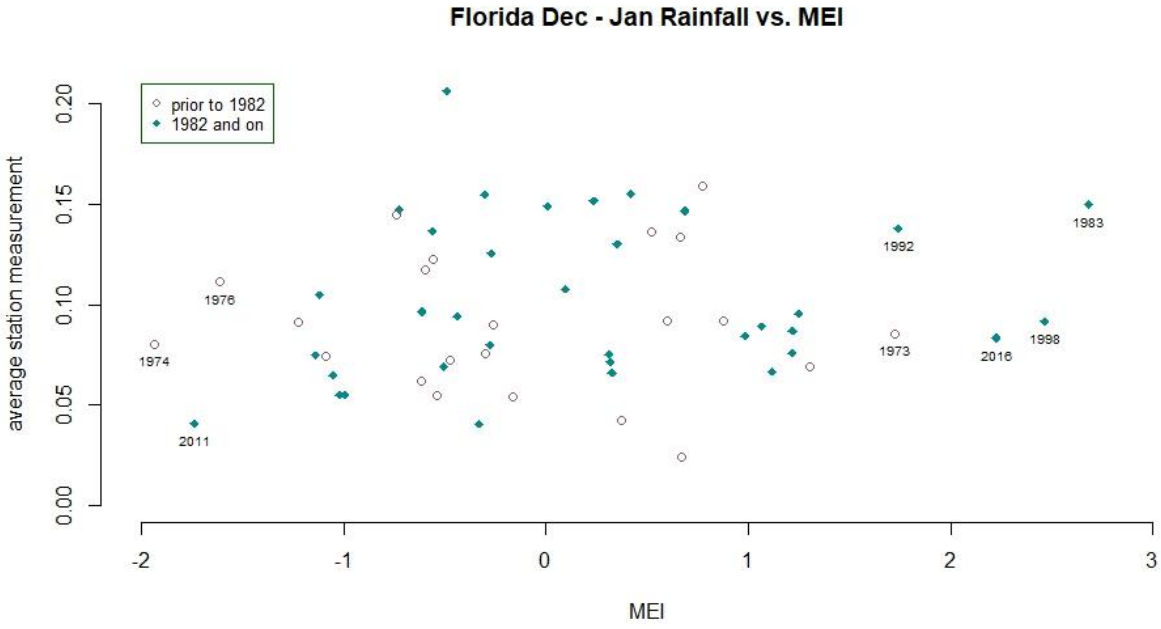


Figure 13. Florida rainfall (PRCP) in December and January plotted against the Multivariate ENSO Index (MEI) appears random.

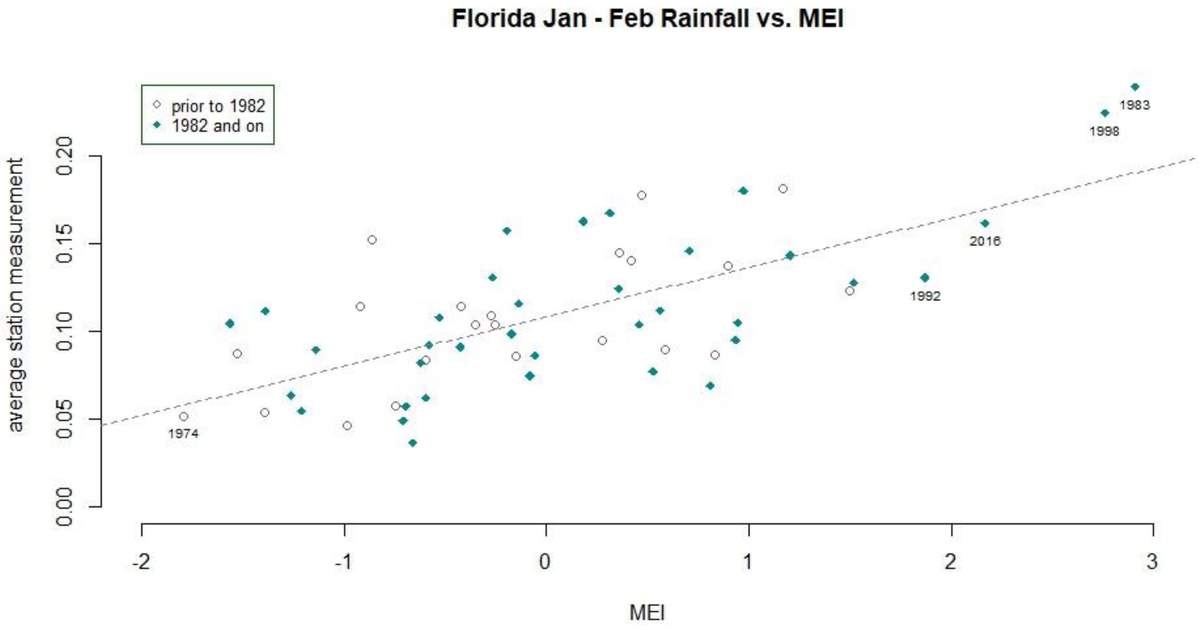


Figure 14. Florida rainfall (PRCP) in January and February plotted against the Multivariate ENSO Index (MEI) reveals a positive correlation.

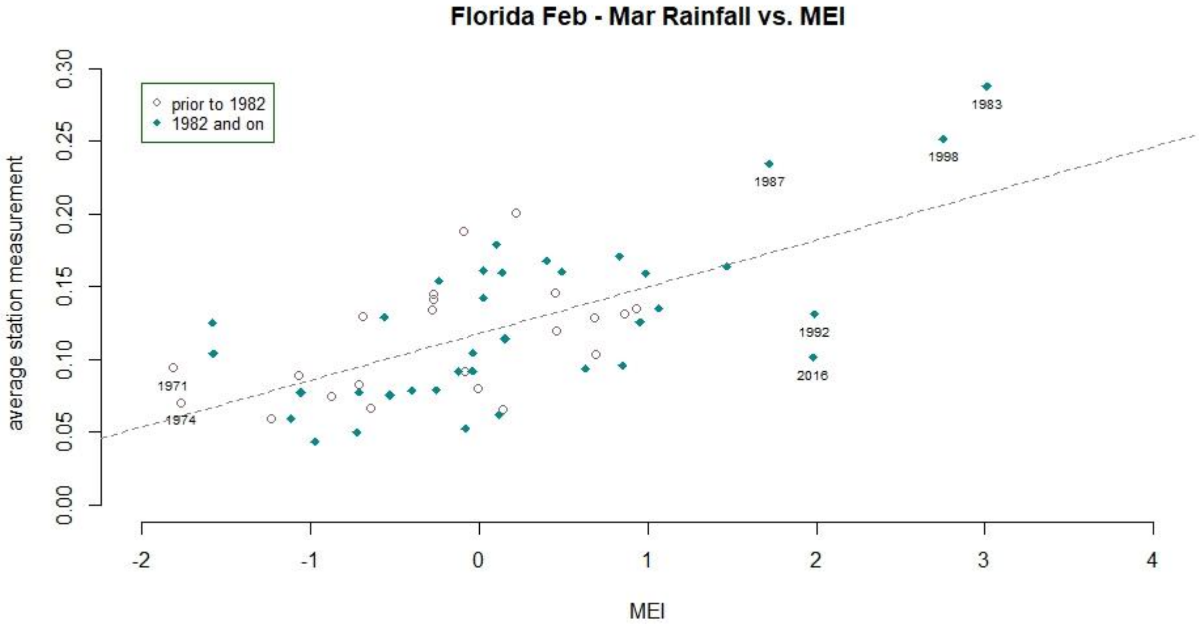


Figure 15. Florida rainfall (PRCP) in February and March plotted against the Multivariate ENSO Index (MEI) reveals a positive correlation.

4. CONCLUSION

With an uncertain future of weather extremes, one might only expect a deluge in climate stances. A detailed raw data source for weather records in GHNC-D can bring some tangibility, at least to the past, to establish a more concrete understanding of the elusive phenomenon of weather. A revised viewpoint would neither presume upward trends in storm losses nor simply level losses to present conditions. Instead the physical process of heat and motion in cycles and patterns, on many scales, might link weather to losses through thresholds. A closer look at distributions and shifts in weather occurring near damageability thresholds might allow losses to definitively enter the climate formula.

If human activity drives any part of climate change, the next technological advancements might be designed to evaluate and financially prepare for the outcome. The objective is to not only use the newest tools to the greatest advantage, but to continually expand our capabilities towards progress, which may include contributions toward an accurate, consistent bank of data with enough stability to distinguish amplitudes, durations and interactions inherent in natural cyclical 'climate change.'

The growing attention to climate as it affects insurance loss may be a calling for actuaries to uncover the hidden message of the meteorological files. The trends in technologies may finally bring the sophisticated topic of climate "down to earth."

Acknowledgement

Many thanks for numerous quality resources provided by the National Oceanic and Atmospheric Administration and to the diverse reviewers including Alp Can FCAS MAAA (actuarial), Dr. Matt Menne (GHCN-D data), Jo Himes (oceanography) and AT Adamack (code).

Supplementary Material

Code and code description.

5. REFERENCES

About Climate Normals. National Oceanic and Atmospheric Administration (NOAA). National Weather Service. Retrieved from www.weather.gov/grr/climatenormals

Ahrens, C. Donald. (2013) *Meteorology Today: An Introduction to Weather, Climate, and the Environment*. Brooks/Cole, Cengage Learning.

Chang, Winston. (2013) *R Graphics Cookbook: Practical Recipes for Visualizing Data*. O'Reilly Media, Inc.

Climate Indices: Monthly Atmospheric and Ocean Time Series. National Oceanic and Atmospheric Administration (NOAA). Earth System Research Library. Physical Sciences Division. Retrieved from www.esrl.noaa.gov/psd/data/climateindices/list/

Download Climate Timeseries. National Oceanic and Atmospheric Administration (NOAA). Global Climate Observing System (GOCS). Atmospheric Observation Panel for Climate (AOPC)/Ocean Observations Panel for Climate (OOPC), Working Group on Surface Pressure. Retrieved from www.esrl.noaa.gov/psd/gcos_wgsp/Timeseries

Earth's Rotation. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Earth%27s_rotation/

Global Historical Climate Network Daily - Methods. National Oceanic and Atmospheric Administration (NOAA). National Centers for Environmental Information (NCEI). Retrieved from www.ncdc.noaa.gov/ghcn-daily-methods

GHCN Monthly. National Oceanic and Atmospheric Administration (NOAA). National Centers for Environmental Information (NCEI). Retrieved from www.ncdc.noaa.gov/ghcnm/

ICAT Damage Estimator. Retrieved from www.icatdamageestimator.com

List of Costliest Atlantic Hurricanes. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/ List_of_costliest_Atlantic_hurricanes

Monthly Atmospheric and SST Indices. National Ocean and Atmospheric Administration (NOAA). National Weather Service. Retrieved from www.cpc.ncep.noaa.gov/data/indices

Orlowski, J. (Director) 2012 *Chasing Ice* [Documentary, Biography] United States: Diamond Docs.

Roger A. Pielke, Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Roger_A._Pielke

Rohli, Robert V. and Vega, Anthony J. (2012) *Climatology*, Second Edition. Jones & Bartlett Learning.

Sarachik, Edward S. and Cane, Mark A. (2010) *The El Niño Southern Oscillation Phenomenon*. Cambridge University Press.

Taylor, Arnold. (2011) *The Dance of Air and Sea: How Oceans, Weather, and Life Link Together*. Oxford University Press.

The Comprehensive R Archive Network. The R Foundation for Statistical Computing. Retrieved from https://cran.r-project.org/

Tropical Weather: *the El Niño Southern Oscillation (ENSO).* National Oceanic and Atmospheric Administration (NOAA). National Weather Service. JetStream - An Online School for Weather. Retrieved from www.srh.noaa.gov/jetstream/tropics/enso_impacts.html

Gwendolyn Anderson ACAS MAAA is an actuary who grew up in a solar (glass) house beside the epicenter of the Loma Prieta earthquake and later achieved an international position in the New York City financial district one month before the 9/11/01 terrorist attacks. Her background is primarily commercial lines pricing, and includes Atlantic hurricanes and financial and predictive models. Off hours she might be found explaining policy coverage in Spanish at a tornado site or exploring earth-friendly technologies at BioGrasse in France. She recently earned a master's degree in financial mathematics focusing on climatology and meteorology, and is especially interested in areas of perplexing logic, unusual probabilities and high risk.

Meteorology for Actuaries – Part 2

Climate and the El Niño Southern Oscillation

Code and Code Description

(February 2018)

5.0 Set up in R

To begin, copy and paste the code into an *.R script file. The code follows the descriptions. If R GUI and R Studio have not yet been installed, instructional videos are available on youtube. After copying code into the *.R script, single quotes may need to be replaced with properly formatted quotes (use <ctrl>-f to find and replace single quotes.) Before running the code, directory paths must be specified and inputs copied into *.csv files.

Directories and Inputs

The paths to three directories are to be specified in the code where R can locate the initial input files and write output files. The input files for this example will be in *.csv (Comma Delimited) and need to be saved to the directory folders named in the code. The files listed below with the directories are the files to be used in the code examples. The code can be modified to run fewer or more years of zipped *.gz files or to read different base inputs. If expanding years of input, be aware that daily data figures are unadjusted and years before 1960 will be subject to inconsistencies in measurements. The files saved in the first two directories (I) dirzip and (II) dirbase provide the inputs to produce subsets and summaries, which are written out as more accessible *.csv files to (III) diroutput. The output files will be accessed again to run subsequent code much more efficiently than by unzipping cumbersome *.gz files.

The three directories and contents will be as follows:

```
(I)    dirzip <- "C:/…/WeatherZip"
                                         1960.csv.gz
                                         1961.csv.gz
                                         …
                                         2016.csv.gz
                                         2017.csv.gz
(II)   dirbase <- "C:/…/WeatherBase"
                                         BEST1mo.csv
                                         CostlyStorms.csv
                                         EQSOI.csv
                                         ghcnd-inventory.csv
                                         ghcnd-stations.csv
                                         ghcnd-states.csv
                                         MEI.csv
                                         NinoMonthly.csv
                                         NinoWeekly.csv
                                         ONI.csv
                                         SOI_Anom.csv
                                         TNI.csv
(III)  diroutput <- "C:/…/WeatherData"
```

The first folder (I) dirzip contains the zipped files daily data, which are downloaded from the NOAA GHCN-Daily at this website address:

ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/

```
by_year/                folder of zipped files of daily data by year to download
readme.txt              detailed descriptions of variables and their values
ghcnd-stations.txt      StationID, name, coordinates, elevation, state/province abbreviation
ghcnd-countries.txt     two-character GHCN country and territory codes, and names
ghcnd-states.txt        two-character US states and territories, Canadian provinces
ghcnd-inventory.txt     StationID, coordinates, station start and end years by element
```

Although text files can be read by R, it is more reliable overall to copy and parse the data into excel and save as *.csv files. The example code runs data for the US and Canada ('CA'). The file ghcnd-countries.txt gives two-character country codes and country names that can be used as inputs to modify the example. The file ghcnd-inventory.txt provides basic ranges of years during which stations have recorded data, by weather element; this inventory list is longer than ghcnd-stations.txt which lists each station once only. The code will summarize greater station detail from the weather records to assist in selecting consistent data across years. Note that the zip code field in ghcnd-stations.txt is missing entries so it would not serve for mapping stations to counties.

The fields 'Open' and 'Close' in the ghcnd-inventory.txt file were included in the example code at a later time so as not to be shown in the sample outputs of this paper.

```
-----------------------------------------------------------------------------------------
Table 1.  Sample output from ghcnd-stations.txt saved as *.csv and read by R as a data table.


     StationID      lat       lon       elev   St    Name                    GSNFlag    zip
1:  ACW00011604    17.1167   -61.7833   10.1   NA    ST JOHNS COOLIDGE FLD     NA       NA
2:  ACW00011647    17.1333   -61.7833   19.2   NA    ST JOHNS                  NA       NA
3:  AE000041196    25.333     55.517    34     NA    SHARJAH INTER. AIRP       NA       41196
4:  AEM00041194    25.255     55.364    10.4   NA    DUBAI INTL                NA       41194
5:  AEM00041217    24.433     54.651    26.8   NA    ABU DHABI INTL            NA       41217
---
*:  ZI000067969    -21.05     29.367     861   NA    WEST NICHOLSON            NA       67969
*:  ZI000067975    -20.067    30.867    1095   NA    MASVINGO                  NA       67975
*:  ZI000067977    -21.017    31.583     430   NA    BUFFALO RANGE             NA       67977
*:  ZI000067983    -20.2      32.616    1132   NA    CHIPINGE                  NA       67983
*:  ZI000067991    -22.217    30         457   NA    BEITBRIDGE                NA       67991

*  column numbers not shown (104122 – 104126)
-----------------------------------------------------------------------------------------
StationID     station identification number
lat           latitude coordinate of station location
lon           longitude coordinate of station location
elev          elevation of station location
St            state or province two-character abbreviation
Name          station name
GSNFlag       (see readme.txt for details)
zip           zip code of station location
```

```
-----------------------------------------------------------------------------------------
Table 2.  Sample output from ghcnd-inventory.txt saved as *.csv and read by R as a data table.

            StationID       lat        lon         elem     Open     Close
 1:         ACW00011604      17.1167   -61.7833     TMAX     1949     1949
 2:         ACW00011604      17.1167   -61.7833     TMIN     1949     1949
 3:         ACW00011604      17.1167   -61.7833     PRCP     1949     1949
 4:         ACW00011604      17.1167   -61.7833     SNOW     1949     1949
 5:         ACW00011604      17.1167   -61.7833     SNWD     1949     1949
 ---
 596841:    ZI000067983    -20.2        32.616      PRCP     1951     2017
 596842:    ZI000067983    -20.2        32.616      TAVG     1962     2017
 596843:    ZI000067991    -22.217      30          TMAX     1951     1990
 596844:    ZI000067991    -22.217      30          TMIN     1951     1990
 596845:    ZI000067991    -22.217      30          PRCP     1951     1990
-----------------------------------------------------------------------------------------
Open          first year the station recorded data for the weather element specified
Close         final year the station recorded data for the weather element specified
```

```
-----------------------------------------------------------------------------------------
Table 3. Sample output from ghcnd-states.txt (left) and ghcnd-countries.txt (right) saved as
*.csv files and read by R as data tables.

                                   ||
       St      Name                ||             loc     CountryName
   1:    AB      ALBERTA             ||      1:      AC      Antigua and Barbuda
   2:    AK      ALASKA              ||      2:      AE      United Arab Emirates
   3:    AL      ALABAMA             ||      3:      AF      Afghanistan
   4:    AR      ARKANSAS            ||      4:      AG      Algeria
   5:    AS      AMERICAN SAMOA      ||      5:      AJ      Azerbaijan
   ---                             ||      ---
   70:   WA     WASHINGTON          ||      214:   WI      Western Sahara
   71:   WI     WISCONSIN           ||      215:   WQ      Wake Island [United States]
   72:   WV     WEST VIRGINIA       ||      216:   WZ      Swaziland
   73:   WY     WYOMING             ||      217:   ZA      Zambia
   74:   YT     YUKON TERRITORY     ||      218:   ZI      Zimbabwe
-----------------------------------------------------------------------------------------
loc           two-character abbreviation for country or territory
```

The file IndexMonthly.csv is created by code, combining various monthly indices that have been accessed separately from online sources and saved into *.csv files. ENSO indices used in the sample code can be copied from these sources:

Website Address	Indices (format: Wide or Long)	from
(I) www.cpc.ncep.noaa.gov/data/indices/	Niño, ONI (L); SOI, EQSOI (W)	1950's
	Niño Weekly (L)	1990's
(II) www.esrl.noaa.gov/psd/gcos_wgsp/Timeseries/	Niño, SOI (W)	1870's
(III) www.esrl.noaa.gov/psd/data/climateindices/list/	Niño, ONI SOI, TNI, BEST, MEI (W)	1950's
(IV) www.esrl.noaa.gov/psd/enso/mei/table.html	MEI (W)	1950's

The first online resource (I) includes all of the monthly Niño indices and anomalies in one file, dating from the 1950's (ERSST monthly). Weekly Niño indices and anomalies are also available in one file, although only from the 1990's (OISST weekly). The ONI is given in a separate file (ERSST seasonal) also from the 1950's. These indices are given in a "long" format, indicated above as (L), where months are stacked in one column. The SOI and EQSOI are each given in separate files from the 1950's in a "wide" format (W) where each month is in a separate column. In R code, the "wide" format can be converted to "long," or vice versa, using package 'tidyr' functions (i.e. gather() and spread()).

The second online resource (II) provides a number of climate indices, including each of the Niño indices given separately in "wide" format from 1870, and each anomaly separately also. The SOI is similarly given in "wide" format monthly back to 1866. The older years may be of limited value for comparison against inconsistent weather data. The third online resource (III) also contains numerous climate indices, and is a source for the Trans-Niño Index (TNI) in wide format from the 1950's. The fourth resource (IV) is the direct site for the Multivariate ENSO Index (MEI).

Running Code

To run one or multiple lines of code, highlight the code and press <ctrl>-r. To run one line of code, alternatively, place the cursor at the line and press <ctrl>-<enter>. To run a 'for loop' highlight the entire loop from 'for' to the end bracket '}' and press <ctrl>-r. Comment lines begin with '#' and will not run.

Packages

```
data.table    functions run faster than base R code
                   rbindlist() to combine years of weather data frames from a list
                   setnames() to update column headers
```

```
tidyverse      a set of packages for organizing data
                    package 'readr' to unzip *.gz files
                    package 'dplyr' to calculate statistics
                    package 'ggplot2' to map choropleths
               package 'tidyr' functions to convert formats between wide and long
lubridate      functions to calculate number of days in month, leap years, etc.
```

Tutorials are available online for instruction on utilizing the data.table functions advantageously.

Common Errors

Because the daily data is voluminous, errors encountered running code may involve space and memory. "Error: cannot allocate vector of size _ Mb" may occur if many large data sets are stored in the environment. The command ls() can be used to view current data sets, and rm() can be used to remove a data set specified within the parentheses. To free memory, the computer can be completely shut down and restarted without opening programs other than R. If a 'for loop' stops prior to completion, the command ls() can be used to identify the latest data set so the code can be continued from that point inside the brackets; a shorter loop can then be defined based on the remaining years or elements.

5.1 Code 1: Weather Daily Loop

```
---------------------------------------------------------------------------------------------
Table 4. Code 1 Sample Output.  Weather Daily Loop.  Precipitation, US and Canada.

   StationID   date     elem VAL  MFlag QFlag SFlag Time loc year month monthday VAL_US
1: CA001010720 19600101 PRCP 0     -     -      C    -   CA  1960  1      101     0
2: CA001010720 19600102 PRCP 25    -     -      C    -   CA  1960  1      102     0.098425
3: CA001010720 19600103 PRCP 0     T     -      C    -   CA  1960  1      103     0
4: CA001010720 19600104 PRCP 41    -     -      C    -   CA  1960  1      104     0.161417
5: CA001010720 19600105 PRCP 257   -     -      C    -   CA  1960  1      105     1.011811
---
*  USW00094967 19601227 PRCP 0     T     -      0    -   US  1960  12     1227    0
*  USW00094967 19601228 PRCP 0     -     -      0    -   US  1960  12     1228    0
*  USW00094967 19601229 PRCP 0     T     -      0    -   US  1960  12     1229    0
*  USW00094967 19601230 PRCP 0     -     -      0    -   US  1960  12     1230    0
*  USW00094967 19601231 PRCP 5     -     -      0    -   US  1960  12     1231    0.019685
---------------------------------------------------------------------------------------------
*  column numbers not shown (3982198 – 3982201)

date           yyyymmdd format
VAL            record in metric system units according to weather element, given in Table 5
MFlag          includes notation 'P' for blank records assumed zero
QFlag          (See readme.txt for details)
SFlag          (See readme.txt for details)
Time           (See readme.txt for details)
year           field created from date
month          field created from date
monthday       field created from date
VAL_US         conversion of VAL field to US Imperial system units, given in Table 5
```

Code 1 is a double loop that unzips the massive meteorological data files containing the daily station detail of weather element measurements. Since unzipping requires the most run time, for each year unzipped the code loops through weather elements to write out to separate *.csv files by element. Wind data is sparse so a few of the GHNC elements are combined in one output file as 'WIND' as a collective term, not a GHNC element. The number of *.csv files to be written out equals the number of years selected in the outer loop times the number of elements selected.

Table 5. Weather elements included in sample code.

FIVE CORE ELEMENTS

Abbr	Element	Unit of Measure	Converted (US)
PRCP	Precipitation	tenths of mm	inches
SNOW	Snowfall	mm	inches
SNWD	Snow depth	mm	inches
TMAX	Maximum temperature	tenths of degrees C	degrees Fahrenheit
TMIN	Minimum temperature	tenths of degrees C	degrees Fahrenheit

WIND elements are coded to include:

Abbr	Element	Unit of Measure	US
AWND	Average daily wind speed	tenths of meters per second	mph
WSF1	Fastest 1-minute wind speed	tenths of meters per second	mph
WSF2	Fastest 2-minute wind speed	tenths of meters per second	mph
WSF5	Fastest 5-second wind speed	tenths of meters per second	mph
WSFG	Peak gust wind speed	tenths of meters per second	mph
WSFI	Highest instantaneous wind speed	tenths of meters per second	mph
WSFM	Fastest mile wind speed	tenths of meters per second	mph

To preserve memory resources, time, and storage, few calculations are made while unzipping. Only five columns are added, the location (country/territory) for selection purposes, a few date fields (year, month, month-day), and the U.S. measurement conversion. The sample countries selected are United States and Canada, which are manageable with 8GB memory. Additional countries or territories may need to be selected separately to avoid errors from inadequate memory; while countries with extremely sparse data may needed to be selected in combination so the code will not stop. Additional weather elements that may be selected are listed in the 'readme.txt' file at the GHCN-D site. The 'readme.txt' file provides descriptions for the fields in the data files represented by the data set column names.

The MFlag notation 'P' in the daily records is quite critical as it represents blank records assumed as zero. This notation applies only to quantity elements like rainfall, and not to continuous measures such as temperature. MFlag also has a notation 'T' that R can mistake for a logical (True/False) causing the 'P' notations to be deleted when writing out to a saved file. The code converts empty cells to dashes which avoids losing data in an unintended format conversion.

Code 3 will adjust counts of zero and blank entries based on the MFLAG 'P' notation, but the notation was not widely used before 1982. The data requires adjustments for unidentified blanks assumed zero prior to 1982, and for the improvement in completion of zero records since 1982.

5.2 Code 2: Detailed Station Inventories

```
--------------------------------------------------------------------------------------------
 Table 6.  Code 2 Sample Output. Detailed Station Inventories.  Precipitation, US and Canada.

          StationID     loc   St    lat      lon     elev    elem    mindate     maxdate
 1:       CA001010066   CA    BC    48.867  -123.3     4     PRCP    19840701    19961129
 2:       CA001010235   CA    BC    48.4    -123.5    17     PRCP    19710601    19950430
 3:       CA001010595   CA    BC    48.583  -123.5    85     PRCP    19610208    19801128
 4:       CA001010720   CA    BC    48.5    -124     351     PRCP    19600101    19710831
 5:       CA001010780   CA    BC    48.333  -123.6    12     PRCP    19600101    19660430
 ---
 56621:   USW00094996   US    NE    40.695   -96.85   NA     PRCP    20020622    20171231
 56622:   USW00096404   US    AK    62.737  -141.2    NA     PRCP    20110926    20171231
 56623:   USW00096406   US    AK    64.501  -154.1    NA     PRCP    20140829    20171231
 56624:   USW00096407   US    AK    66.562  -159      NA     PRCP    20150814    20171231
 56625:   USW00096408   US    AK    63.452  -150.9    NA     PRCP    20150820    20171214

          minyear    minmo    maxyear    maxmo     clsdmbeg    clsdmend    clsdinm    clsdbef
 1:         1984      7        1996       11          0           1           1         182
 2:         1971      6        1995        4          0           0           0         151
 3:         1961      2        1980       11          7           2           9          31
 4:         1960      1        1971        8          0           0           0           0
 5:         1960      1        1966        4          0           0           0           0
 ---
 56621:     2002      6        2017       12         21           0          21         151
 56622:     2011      9        2017       12         25           0          25         243
 56623:     2014      8        2017       12         28           0          28         212
 56624:     2015      8        2017       12         13           0          13         212
 56625:     2015      8        2017       12         19          17          36         212

         clsdaft    clsdfulm    bsyrct    rcyrct     bryrct     bsspan     rcspan     brspan
 1:        31         213         7         6          13          7          6          13
 2:        245        396         8         5          13         20          5          25
 3:        31         62          20        0          20         20          0          20
 4:        122        122         11        0          11         11          0          11
 5:        245        245         6         0           6          6          0           6
 ---
 56621:   0           151         0         16         16          0         16          16
 56622:   0           243         0         7           7          0          7           7
 56623:   0           212         0         4           4          0          4           4
 56624:   0           212         0         3           3          0          3           3
 56625:   0           212         0         3           3          0          3           3
--------------------------------------------------------------------------------------------
StationID      Station identification number from GHCN-D.
loc            two-character country code, the first two digits of the StationID
St             two-character state or province abbreviation where station is located
lat            latitude coordinate of station location
lon            longitude coordinate of station location
elev           elevation of station location
elem           weather element
Open           (not shown) first year of station records from ghcnd-inventory.csv
Close          (not shown) last year of station records from ghcnd-inventory.csv
mindate        earliest date of station record, from years selected to summarize
maxdate        latest date of station record, from years selected to summarize
minyear        year of mindate
minmo          month of mindate
maxyear        year of maxdate
maxmo          month of maxdate
clsdmbeg       count of days the station was not yet operating at beginning of partial month
```

```
clsdmend       count of days the station was no longer operating at end of partial month
clsdinm        total days station did not operate in partial months (clsdmbeg + clsdmend)
clsdbef        count of days in year the station was not yet operating for full months
clsdaft        count of days in year the station no longer was operating for full months
clsdfulm       count of days in year the station was closed for full months (clsdbef + clsdaft)
bsyrct         count of years with observations in the selected base period
rcyrct         count of years with observations in recent years following base period
bryrct         count of years with observations in base and recent years
bsspan         span of base years that fall within the mindate and the maxdate
rcspan         span of years following the base period within the mindate and the maxdate
brspan         span of base and recent years that fall within the mindate and the maxdate
```

Code 2 is an intermediate step to conserve memory. Code 2 reads in the subsets of daily data from Code 1 and produces summaries of station inventories separately for each weather element, for all stations with at least one record in the years of data selected. The focus of the summary is to calculate the minimum and maximum date of record. These dates are compared against the station's opening and closing years given in ghcnd-inventory.txt, to arrive at counts of days the station was not operating in partial months (also full months), for the month (also the year) the station opened or closed. The summary counts years with records for the selected base years and recent years, and also calculates the span of time from the earliest to latest year of record without deducting empty data years. The summary again merges, with data from ghcnd-stations.txt, to list each station's location by coordinates and state. All years from the daily files are combined into one file for each element, so the number of files output is equal to the number of elements selected. If the ghnc-station.txt list is incomplete, the two-character state abbreviation can be found in the StationID for more recent years; but the elevation and coordinates will be missing from final outputs.

5.3 Code 3: Initial Year-Month Summaries

```
-----------------------------------------------------------------------------------------
 Table 7.  Code 3 Sample output. Initial Year-Month Summary.  Precipitation, US and Canada.

          StationID      loc    elem     year   month     VALm     VALm_US     sumVALsqd_US
 1:       CA001010720    CA     PRCP     1960     1       3497     13.7677165    20.0084785
 2:       CA001010720    CA     PRCP     1960     2       4155     16.3582677    29.8169911
 3:       CA001010720    CA     PRCP     1960     3       3839     15.1141732    24.9078523
 4:       CA001010720    CA     PRCP     1960     4       3486     13.7244094    16.6643623
 5:       CA001010720    CA     PRCP     1960     5       2412      9.496063      6.6623473
 ---
 9001954: USW00096408    US     PRCP     2017     8        800      3.1496063     1.1848844
 9001955: USW00096408    US     PRCP     2017     9        623      2.4527559     0.7029729
 9001956: USW00096408    US     PRCP     2017     10       895      3.523622      1.6202957
 9001957: USW00096408    US     PRCP     2017     11       598      2.3543307     0.6302003
 9001958: USW00096408    US     PRCP     2017     12       211      0.8307087     0.1854269

             VALsqm_US     zerrec     recs    zblank    zerobs    obs    daysinmo   misclsd
 1:         189.5500186      10        31       0         10      31       31         0
 2:         267.5929227      12        29       0         12      29       29         0
 3:         228.4382324      11        31       0         11      31       31         0
 4:         188.3594147       9        30       0          9      30       30         0
 5:          90.1752124       6        31       0          6      31       31         0
 ---
 9001954:     9.9200198      9       30      0        9      30      31        1
 9001955:     6.0160115     15       30      0       15      30      30        0
 9001956:    12.4159123     15       31      0       15      31      31        0
 9001957:     5.5428731     17       30      0       17      30      30        0
 9001958:     0.6900769      9       14      0        9      14      31       17
-------------------------------------------------------------------------------------
VALm           monthly sum of daily VAL
VALm_US        monthly sum of daily VAL_US, VAL converted to U.S. unit of measure.
sumVALsqd_US   sum of squared daily values in U.S. units, to compute daily variances
VALsqm_US      squared monthly values in U.S. units, to computer monthly variances
zerrec         count of records with zero values (quantities); zero and below (temperature)
recs           count of all records
zblank         blank records assumed zero and recorded as zero (for quantity elements)
zerobs         net (below) zero observations, after subtracting blanks assumed zero
obs            net observations, after subtracting blanks assumed zero
daysinmo       count of days in calendar month, representing maximum possible records
misclsd        count of days in month with missing records, or days station not open
```

Code 3 is an intermediate step to conserve memory. Code 3 reads in the subsets of daily data from Code 1 and summarizes monthly data. It also provides important counts of records, for use in selecting stations with adequately complete data or for identifying areas in need of adjustments.

A critical threshold is calculated as variable 'zerobs' which for quantities like rainfall counts zero values relating to drought. For temperatures, the critical threshold 'zerobs' represents freezing temperatures, zero degrees Celsius and below.

A critical data adjustment is made to quantity records through a code calculation, deducting blank records assumed zero ('zblank') from zero records ('zerrec') and records ('recs') to arrive at the adjusted count of observations ('obs') and zero observations ('zerobs'). This calculation does not hold for years before 1982 when the MFlag notation 'P' was rarely used to identify blanks assumed zeros.

No adjustment has been made to the raw daily data records either for blank records assumed zero, or to account for the continued improvement in record completion since 1982. Therefore, the GHNC-D data is in need of adjustments for changes in record keeping over time. No other adjustments are made by the example code, besides converting 'zblank' record counts from zero to blank.

A number of adjustments have been made in the monthly summary GHNC-M, available online [www.ncdc.noaa.gov/ghcnm] with references on the 'homogeneity adjustment' that could be considered when analysis necessitates consistency along with the detail of the raw daily data set.

5.4 Code 4: Complete Year and Year-Month Summaries Merged with Station Locations and Inventories

```
-----------------------------------------------------------------------------------
Table 8.  Code 4 Sample output (I).  Complete Year-Month Summary.  Precipitation, US and
Canada

       StationID     loc    St    lat     lon     elev   mindate    maxdate
 1:    CA001010720   CA     BC    48.5    -124    351    19600101   19710831
 2:    CA001010720   CA     BC    48.5    -124    351    19600101   19710831
 3:    CA001010720   CA     BC    48.5    -124    351    19600101   19710831
 4:    CA001010720   CA     BC    48.5    -124    351    19600101   19710831
 5:    CA001010720   CA     BC    48.5    -124    351    19600101   19710831
 6:    CA001010720   CA     BC    48.5    -124    351    19600101   19710831

       elem     year   month     VALm     VALm_US     sumVALsqd_US   VALsqm_US
 1:    PRCP     1960    1        3497     13.768       20.008        189.550019
 2:    PRCP     1960    2        4155     16.358       29.817        267.592923
 3:    PRCP     1960    3        3839     15.114       24.908        228.438232
 4:    PRCP     1960    4        3486     13.724       16.664        188.359415
 5:    PRCP     1960    5        2412      9.4961       6.6623        90.175212
 6:    PRCP     1960    6         535      2.1063       0.7874         4.436496

       zerrec    recs   zblank     zerobs     obs     daysinmo   misinm   clsdinm
 1:      10       31      0          10       31        31         0         0
 2:      12       29      0          12       29        29         0         0
 3:      11       31      0          11       31        31         0         0
 4:       9       30      0           9       30        30         0         0
 5:       6       31      0           6       31        31         0         0
 6:      18       30      0          18       30        30         0         0
-----------------------------------------------------------------------------------
clsdinm     count of days station closed in partial month, sum (clsdmbeg + clsdmend)
misinm      count of missing records in partial month, the difference (misclsd - clsdinm)

-----------------------------------------------------------------------------------
Table 9.  Code 4 Sample output (II).  Complete Year Summary. Precipitation, US and Canada

     StationID      loc    St    lat       lon         elev    mindate   maxdate
 1:  CA001010720    CA     BC    48.5      -124         351    19600101  19710831
 2:  CA001010780    CA     BC    48.3333   -123.633      12    19600101  19660430
 3:  CA001010965    CA     BC    48.5667   -123.433      91    19600801  19700630
 4:  CA001011500    CA     BC    48.9333   -123.75       75    19600101  20171231
 5:  CA001011920    CA     BC    48.5333   -123.367      37    19600101  19700331
 6:  CA001012010    CA     BC    48.7167   -123.55        1    19600101  20010311

      elem    year     VALy     VALy_US   sumVALsqd_US    sumVALsqm_US     VALsqy_US
 1:   PRCP    1960    33091   130.27953    198.905868      1892.87716     16972.7553
 2:   PRCP    1960     9473    37.29528     24.867707       170.58961      1390.9376
 3:   PRCP    1960     3842    15.12598      8.665075        63.39429       228.7954
 4:   PRCP    1960    11755    46.27953     39.69544        259.06651      2141.7947
 5:   PRCP    1960     7778    30.62205     16.595573       107.77906       937.7098
 6:   PRCP    1960     9984    39.30709     29.90204        184.49805      1545.0471

      zerrec      recs      zblank    zerobs      obs      daysum    daysinyr     monthct
 1:     167       366          0       167       366        366        366          12
 2:     208       366          0       208       366        366        366          12
 3:      82       148          0        82       148        153        366           5
 4:     210       366          0       210       366        366        366          12
 5:     223       364          0       223       364        366        366          12
 6:     208       366          0       208       366        366        366          12
```

```
     clsdinm   clsdfulm  clsdall   misinm   misfulm   misall
1:        0         0        0        0        0        0
2:        0         0        0        0        0        0
3:        0       213      213        5        0        5
4:        0         0        0        0        0        0
5:        0         0        0        2        0        2
6:        0         0        0        0        0        0

     bsyrct    rcyrct    bryrct    bsspan    rcspan  brspan
1:       11         0        11        11        0       11
2:        6         0         6         6        0        6
3:       10         0        10        10        0       10
4:       17        27        44        30       27       57
5:       10         0        10        10        0       10
6:       28        11        39        30       11       41
```

Code 4 combines the output of Code 2 and 3, the detailed station inventory and the initial year-month summary. The code produces a year-month summary with station detail and similarly a detailed yearly summary. The year-month summary gains a more complete record count from the station detail; a month where records are missing or stations are not operating is merged to match the partial month of the data summary. The days in a month the station is not open are distinguished from the days the station is open but records are missing. The yearly summary further includes day counts for full months in the year where all records are missing or the station is closed.

5.5 Code 5: Missing Records by Year

```
---------------------------------------------------------------------------------------------
 Table 10.  Code 5 Sample Output.  Missing Records by Year.  All sample weather elements, US
 and Canada.

          loc      elem       year      zerrec         recs    zblank    zerobs        obs
1:        US       AWND       1982           0          243         0         0        243
2:        US       AWND       1984          63       102531         0        63     102531
3:        US       AWND       1985         111       101691         0       111     101691
4:        US       AWND       1986          57       104826         0        57     104826
5:        US       AWND       1987          46       116725         0        46     116725
---
881:      CA       WSFG       1969           0          364         0         0        364
882:      CA       WSFG       1970           0          136         0         0        136
883:      CA       WSFG       2015       70208       175879         0     70208     175879
884:      CA       WSFG       2016       95246       235934         0     95246     235934
885:      CA       WSFG       2017       74654       196613         0     74654     196613

     stndays stndysinyr  stnmos  clsdinm  clsdfulm   clsdall   misinm   misfulm   misall   stnct
1:       243        365       8        0         0         0        0       122      122       1
2:    102968     105042    3376       16       640       656      421      1434     1855     287
3:    102074     102930    3356       17        92       109      366       764     1130     282
4:    105445     119355    3467        7     11421     11428      612      2489     3101     327
5:    117619     118260    3867        1       579       580      893        62      955     324
---
881:     365        365      12        0         0         0        1         0        1       1
882:     151        365       5        7       214       221        8         0        8       1
883:  180861     247470    5919      536     64712     65248     4446      1897     6343     678
884:  240462     247416    7884      379      3968      4347     4149      2986     7135     676
885:  200453     239075    6597      779     36251     37030     3061      2371     5432     655
```

```
     stndays  stndysinyr  stnmos  clsdinm  clsdfulm  clsdall  misinm   misfulm   misall
 1:      243         365       8        0         0        0       0       122      122
 2:   102968      105042    3376       16       640      656     421      1434     1855
 3:   102074      102930    3356       17        92      109     366       764     1130
 4:   105445      119355    3467        7     11421    11428     612      2489     3101
 5:   117619      118260    3867        1       579      580     893        62      955
 ---
 881:    365         365      12        0         0        0       1         0        1
 882:    151         365       5        7       214      221       8         0        8
 883: 180861      247470    5919      536     64712    65248    4446      1897     6343
 884: 240462      247416    7884      379      3968     4347    4149      2986     7135
 885: 200453      239075    6597      779     36251    37030    3061      2371     5432

        stnct    pctobsyr    pctmisyr    pctclsdyr  pctzblkyr   pctzeroyr     pctzerobs
 1:         1    0.665753    0.334247    0               0        0             0
 2:       287    0.976095    0.01766     0.006245        0        0.0006        0.000614
 3:       282    0.987963    0.010978    0.001059        0        0.001078      0.001092
 4:       327    0.878271    0.025981    0.095748        0        0.000478      0.000544
 5:       324    0.98702     0.008075    0.004904        0        0.000389      0.000394
 ---
 881:       1    0.99726     0.00274     0               0        0             0
 882:       1    0.372603    0.021918    0.605479        0        0             0
 883:     678    0.710708    0.025631    0.26366         0        0.283703      0.399184
 884:     676    0.953592    0.028838    0.01757         0        0.384963      0.403698
 885:     655    0.822391    0.022721    0.154889        0        0.312262      0.3797
-------------------------------------------------------------------------------------
```

5.6 Code 6: Multiple Month Indices

```
-------------------------------------------------------------------------------------
Table 11.  Code 6 Sample output.  Multiple month indices (Bimonthly).  Precipitation, US and
Canada.

          StationID     loc   St    lat        lon       elev   mindate   maxdate    elem
1:        CA001010066   CA    BC    48.8667   -123.283     4    19840701  19961129   PRCP
2:        CA001010066   CA    BC    48.8667   -123.283     4    19840701  19961129   PRCP
3:        CA001010066   CA    BC    48.8667   -123.283     4    19840701  19961129   PRCP
4:        CA001010066   CA    BC    48.8667   -123.283     4    19840701  19961129   PRCP
5:        CA001010066   CA    BC    48.8667   -123.283     4    19840701  19961129   PRCP
---
9184276:  USW00096406   US    AK    64.5014   -154.13     NA    20140829  20171231   PRCP
9184277:  USW00096407   US    AK    66.562    -159.004    NA    20150814  20171231   PRCP
9184278:  USW00096407   US    AK    66.562    -159.004    NA    20150814  20171231   PRCP
9184279:  USW00096408   US    AK    63.4519   -150.875    NA    20150820  20171214   PRCP
9184280:  USW00096408   US    AK    63.4519   -150.875    NA    20150820  20171214   PRCP

          year    ord    bimo     MEI       VAL2m     VAL2m_US    sumVALsqd_US   sumVALsqm_US
 1:       1985    1      JanFeb   -0.595    60        0.2362      0.042904086    5.58E-02
 2:       1986    1      JanFeb   -0.195    2822      11.11       9.504433009    6.45E+01
 3:       1987    1      JanFeb   1.205     1332      5.2441      1.477462955    1.57E+01
 4:       1988    1      JanFeb   0.706     870       3.4252      1.73507347     8.60E+00
 5:       1989    1      JanFeb   -1.262    1254      4.937       1.710893422    1.40E+01
 ---
 9184276: 2016    12     DecJan   2.227     5         0.0197      0.000387501    3.88E-04
 9184277: 2015    12     DecJan   0.42      33        0.1299      0.008137516    1.69E-02
 9184278: 2016    12     DecJan   2.227     111       0.437       0.034425569    9.56E-02
 9184279: 2015    12     DecJan   0.42      0         0           0              0.00E+00
 9184280: 2016    12     DecJan   2.227     470       1.8504      0.351695703    1.72E+00
```

```
         zerrec    recs  zblank  zerobs    obs    daysum    misinm  clsdinm   monthct
 1:          23      25      0     23      25        31         6       0         1
 2:          27      58      0     27      58        59         1       0         2
 3:          25      55      0     25      55        59         4       0         2
 4:          36      51      0     36      51        60         9       0         2
 5:          24      49      0     24      49        59        10       0         2
 ---
 9184276:     0       1      0      0       1        31        30       0         1
 9184277:    19      23      0     19      23        62        39       0         2
 9184278:    16      24      0     16      24        62        38       0         2
 9184279:     1       1      0      1       1        31        30       0         1
 9184280:    39      60      0     39      60        62         2       0         2

------------------------------------------------------------------------------------------
ord        sorting order for multiple month time period based on first month of period
bimo       bimonthly time period for MEI (JanFeb, FebMar, …, NovDec)
trimo      (not shown) trimonthly time period for ONI (JFM, FMA, …, DJF)
MEI        bimonthly Multi-Variate ENSO Index
ONI        (not shown) trimonthly Ocean Niño Index
misinm     sum of missing records for combined months of bimo or trimo period
clsdinm    sum of days station closed for combined months of bimo or trimo period
monthct    count of months with records for bimo or trimo period
```

Some ENSO indices are based on multiple months of data. Bimonthly sums are merged with the Multivariate ENSO Index (MEI) and trimonthly sums are merged with the Oceanic Niño Index (ONI). A count of months is made to indicate at least one record was present in each month. The detail of observations by month can be viewed in the year-month summary. Stations may be selected having observations near to 60 (bimonthly) or 90 (trimonthly) or a minimum monthly observation can be established utilizing the year-month summary for station selection.

5.7 Code 7: Weekly Niño Indices

```
------------------------------------------------------------------------------------------
Table 12.  Code 7 Sample output.  Weekly Niño indices.  Precipitation, US and Canada.

       StationID     loc   elem   St    lat       lon      elev   weekno   yrgrp   ctrweek
 1:    CA001010066   CA    PRCP   BC    48.8667  -123.283     4   1566     1991    1/2/1991
 2:    CA001010960   CA    PRCP   BC    48.6     -123.467    38   1566     1991    1/2/1991
 3:    CA001011467   CA    PRCP   BC    48.5833  -123.417    53   1566     1991    1/2/1991
 4:    CA0010114F6   CA    PRCP   BC    48.5667  -123.4      38   1566     1991    1/2/1991
 5:    CA001011743   CA    PRCP   BC    48.6833  -123.6      99   1566     1991    1/2/1991
 ---
 544315: USW00094911   US    PRCP   SD    42.8783   -97.3633   NA   1617     1991    12/25/1991
 544316: USW00094918   US    PRCP   NE    41.3536   -96.0233   NA   1617     1991    12/25/1991
 544317: USW00094931   US    PRCP   MN    47.3864   -92.8389   NA   1617     1991    12/25/1991
 544318: USW00094957   US    PRCP   NE    40.0803   -95.5919   NA   1617     1991    12/25/1991
 544319: USW00094967   US    PRCP   MN    46.9006   -95.0678   NA   1617     1991    12/25/1991

        VALw    VALw_US     sumVALsqd_US  VALsqw_US   zerrec  recs  zblank  zerobs  obs  misclsd
 1:        0    0           0             0            6      6      0        6     6      8
 2:       85    0.3346457   0.04738359    0.07168764   4      7      0        4     7      7
 3:      140    0.5511811   0.13751628    0.2015004    4      7      0        4     7      7
 4:       96    0.3779528   0.06782814    0.08642817   4      7      0        4     7      7
 5:       54    0.2125984   0.02287805    0.02287805   4      6      0        4     6      8
 ---
 544315:   0     0            0            0            7     7     7      0    0     7
 544316:  145    0.5708661    0.31268213   0.32588815   5     7     0      5    7     0
 544317:   0     0            0            0            7     7     0      7    7     0
 544318:  117    0.4606299    0.2015779    0.21217992   5     7     0      5    7     0
 544319:   0     0            0            0            7     7     0      7    7     0

        Nino12Ind  Nino12Anom  Nino3Ind   Nino3Anom  Nino34Ind  Nino34Anom  Nino4Ind  Nino4Anom
 1:        23.2        0.5         25.3        0.1         26.9        0.4         28.9     0.5
 2:        23.2        0.5         25.3        0.1         26.9        0.4         28.9     0.5
 3:        23.2        0.5         25.3        0.1         26.9        0.4         28.9     0.5
 4:        23.2        0.5         25.3        0.1         26.9        0.4         28.9     0.5
 5:        23.2        0.5         25.3        0.1         26.9        0.4         28.9     0.5
 ---
 544315:   23.6        0.3         26.7        1.4         28.5        1.9         29.6     1.2
 544316:   23.6        0.3         26.7        1.4         28.5        1.9         29.6     1.2
 544317:   23.6        0.3         26.7        1.4         28.5        1.9         29.6     1.2
 544318:   23.6        0.3         26.7        1.4         28.5        1.9         29.6     1.2
 544319:   23.6        0.3         26.7        1.4         28.5        1.9         29.6     1.2
-----------------------------------------------------------------------------------------------
VALw             weekly sum of daily VAL
VALw_US          weekly sum of VAL_US, VAL converted to U.S. unit of measure.
sumVALsqd_US     weekly sum of squared daily values in U.S. units, to compute daily variances
```

Niño indices are available monthly at least since 1950, and weekly since 1990. The weekly indices give more specific information about the ENSO phase that might be relevant to the timing of losses. The code produces weekly summaries by element according to the weekly groupings of the Niño indices, which begin on 12/31/1989 and are always seven days in length. The result of these divisions will be the same beginning on 1/1/1961. As these weeks will shift through years, they will not provide comparison periods year by year. An alternative numbering scheme is provided at the end of the code labels an eight day week at the end of each year and an eight day week with every leap day. The eighth days weeks can be excluded from assignment if desired. This alternative provides a comparative basis among years but does not match the weekly Niño indices time periods so the index values might be interpolated. The code counts the number of observations by week so that the level of completeness can be determined and used in station selection.

```
-----------------------------------------------------------------------------------------------
Table 13.  Weekly Niño numbering scheme (Table 13-A, left) with seven days per week, and
alternative scheme for comparisons among years (Table 13-B, right) with eight day weeks at the
end of each year and with each leap day.

      weekno  date        ctrweek     yrgrp     ||         weekno   mdchar   ctrweek
  1      1    19610101    1961-01-04  1961      ||    1      1       101     1961-01-04
  2      1    19610102    1961-01-04  1961      ||    2      1       102     1961-01-04
  3      1    19610103    1961-01-04  1961      ||    3      1       103     1961-01-04
  4      1    19610104    1961-01-04  1961      ||    4      1       104     1961-01-04
  5      1    19610105    1961-01-04  1961      ||    5      1       105     1961-01-04
  6      1    19610106    1961-01-04  1961      ||    6      1       106     1961-01-04
  7      1    19610107    1961-01-04  1961      ||    7      1       107     1961-01-04
  8      2    19610108    1961-01-11  1961      ||    8      2       108     1961-01-11
  9      2    19610109    1961-01-11  1961      ||    9      2       109     1961-01-11
  10     2    19610110    1961-01-11  1961      ||    10     2       110     1961-01-11
  11     2    19610111    1961-01-11  1961      ||    11     2       111     1961-01-11
  12     2    19610112    1961-01-11  1961      ||    12     2       112     1961-01-11
-----------------------------------------------------------------------------------------------
```

```
weekno     (A) each week is given its own number; (B) weeks 1 – 52 each year
date       (A) always 7 days in week, field to merge with element or loss   records
ctrweek    weeks are identified by central day of the week.  (A) matches Niño indices.
yrgrp      indicates output *.csv file particularly for weeks overlapping two years
mdchar     (B) month-day will correspond to the same week number in every year.
```

5.8 Code 8 State Summaries Visual Analysis with Choropleth Maps

```
--------------------------------------------------------------------------------------
 Table 14. Code 8 sample of summary by state, with anomalies to be plotted on choropleth maps.
 Precipitation, US by state.

         loc   St    elem   year    VALy      VALy_US     obs     stnct   VALsqSt_US
 1:      US    AL    PRCP   1960    487419    1918.9724   14162   39      3682455.2
 2:      US    AZ    PRCP   1960    159360    627.4016    21350   60      393632.7
 3:      US    AR    PRCP   1960    796665    3136.4764   24788   70      9837484.1
 4:      US    CA    PRCP   1960    1083230   4264.685    75153   208     18187538.5
 5:      US    CO    PRCP   1960    246999    972.437     26517   74      945633.7
 6:      US    CT    PRCP   2016    74192     292.0945    2921    8       85319.19

        bsstnobs    bsstnyrs   bsmean        bsstdev      rm.na    anom
 1:     396676      1200       2239.84       319.3026     TRUE     -1.0049014
 2:     544398      1800       1050.5949     1052.0216    TRUE     -0.4022668
 3:     691713      2130       3467.5654     556.574      TRUE     -0.5948697
 4:     1874258     6270       4607.6652     1421.5972    TRUE     -0.241264
 5:     722090      2280       1122.9764     165.4945     TRUE     -0.9096334
 6:     79998       240        366.5045      61.08135     TRUE     -1.2182111

        Name           STDEVgrp      colreg
 1:     alabama        neg1.5        wheat2
 2:     arizona        neg0.5        lightyellow1
 3:     arkansas       neg0.5        lightyellow1
 4:     california     pos0.50       lightcyan1
 5:     colorado       neg0.5        lightyellow1
 6:     connecticut    neg1.5        wheat2
--------------------------------------------------------------------------------------
```

Code 8 summarizes by state, greatly reducing the data set size, and from the state data creates choropleth maps – maps that color code a region according to a selected data field. Two different packages are used to create choropleth maps, maps that color code a selected data field by region. The packages are 'maps' and 'ggplot2.' In this case the regions are states, and the example data field is the anomalous rainfall in the year mapped as compared against the base climate period. Choropleths of anomalies centered at zero are not as immediately plotted in R because the choropleth packages are primarily designed for positive values such as census data.

Package 'maps' does not plot Alaska or Hawaii with the mainland United States, but the code is simple and straightforward to use for creating a choropleth, and individual states can still be plotted separately. The example choropleth will be created 'from scratch,' meaning a column of colors will be defined in the data set corresponding to the anomaly for each state. In this case, the positive anomalies will be assigned deepening shades of blue to indicate more rainfall at a glance, while the negative anomalies will be assigned deepening shades of tan and brown to indicate dryness. Points are easily plotted on the choropleth to show the location of each station selected for the data underlying the choropleth.

Package ‘ggplot2’ has an advantage over ‘maps’ in its compatibility with the package ‘fiftystater’ which plots insets for Alaska and Hawaii. This package also allows a midpoint to be defined at zero, and will either assign automatic base colors or else will assign grades of colors based on selections made for the low, midpoint, and high values. Selecting a midpoint color with contrast to the low and high value colors will produce a choropleth that is easily interpreted.

To avoid plot errors and formatting glitches, be sure to expand the plot region large enough for map to fit, and for the legend to fit to the side of the map.

The code uses package ‘dplyr’ to produce statistics for the base climate period, but also provides the formulas for calculating anomalies directly. Note that if the package ‘plyr’ is loaded in R, then ‘dplyr’ will not complete the calculations unless ‘plyr’ is detached.

```
-------------------------------------------------------------------------------------
Table 15.  Code 8 sample calculation of base year statistics, mean and standard deviation.
Precipitation, US by state.

       loc      St     elem    bsstnobs   bsstnyrs    bsmean      bsstdev      rm.na
       <chr>    <chr>  <chr>   <int>      <dbl>       <dbl>       <dbl>        <lgl>
1      US       AK     PRCP    348076     990         1354.76     118.8982     TRUE
2      US       AL     PRCP    396676     1200        2239.84     319.3026     TRUE
3      US       AR     PRCP    691713     2130        3467.565    556.574      TRUE
4      US       AZ     PRCP    544398     1800        1050.595    1052.0216    TRUE
5      US       CA     PRCP    1874258    6270        4607.665    1421.5972    TRUE
6      US       CO     PRCP    722090     2280        1122.976    165.4945     TRUE
-------------------------------------------------------------------------------------
```

5.9 Code 9 Combine Monthly Indices

```
-------------------------------------------------------------------------------------
Table 16.  Code 9 Sample Output. Monthly ENSO indices combined into a single file
‘IndexMonthly.csv’ with all indices converted to ‘long’ format.

   year month Nino12 Anom12 Nino3 Anom3 Nino4 Anom4 Nino34 Anom34   SOI EQSOI  BEST    TNI
  <int> <int>  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl>
1  1951     1  24.11  -0.44 24.79 -0.87 27.21 -1.02  25.24  -1.31   2.5   0.1 -1.13  1.315
2  1951     2  25.19  -0.83 25.65 -0.76 27.09 -1.01  25.71  -1.03  -1.5  -1.4  0.64  0.168
3  1951     3  25.74  -0.68 26.87 -0.28 27.74 -0.47  26.90  -0.33   0.5  -0.2  0.18 -0.027
4  1951     4  25.29  -0.18 27.37 -0.11 28.21 -0.24  27.58  -0.13   1.1   0.2  0.00 -0.655
5  1951     5  24.59   0.33 27.07 -0.09 29.18  0.43  27.92   0.11  -0.9  -0.2 -0.01  0.316
-------------------------------------------------------------------------------------
```

This interim code combines various ENSO index files into one, for convenience, utilizing a common “long” format. The Niño indices are promulgated in the “long” format while other indices are available formatted “wide.” For ENSO indices, the “long” format includes two separate columns for year and month, and one column per index; while the “wide” format includes separate columns for each month of the year. It is most practical to utilize the “long” format for all indices, in order to merge data by year and month, and to label each index as a column header.

5.10 Code 10 Plot ENSO Index Time Series

```
------------------------------------------------------------------------------------
Table 17. Sample output of Multivariate ENSO Index (MEI) for plotting time series.

        bimo        Year     MEI      posM      month
 273    JanFeb      1950    -1.163    FALSE       1
 205    FebMar      1950    -1.312    FALSE       2
 477    MarApr      1950    -1.098    FALSE       3
 1      AprMay      1950    -1.445    FALSE       4
 545    MayJun      1950    -1.376    FALSE       5

------------------------------------------------------------------------------------
```

This code produces time series bar plots of indices from MEI, ONI, and the monthly indices previously combined into one dataset by Code 9. The column "positive" is added to discern index values above and below the x-axis, corresponding to positive and negative ENSO phases color coded in the bar graph. The plots are formatted for anomalies and could be modified to present SST measures

For the monthly ENSO indices, a vector "IndexName" is defined by the column headings for the index values. The ENSO index to be plotted is selected by its number position in the vector. The code replaces the column heading with "PlotIndex" which locates the data to plot. After plotting the column heading is returned to the original ENSO index title. If errors are encountered, reset the column headers to the original headers.

```
------------------------------------------------------------------------------
Table 18.  Column headings for data set 'indices' given by names(indices), before
selecting the index (top) and after, where SelIndex <- 4 (bottom).

  [1] "year"   "month"   "Nino12"  "Anom12"  "Nino3"  "Anom3"     "Nino4"
  [8] "Anom4"  "Nino34"  "Anom34"  "SOI"     "EQSOI"  "BEST"      "TNI"

  [1] "year"   "month"   "Nino12"  "Anom12"  "Nino3"  "PlotIndex" "Nino4"
  [8] "Anom4"  "Nino34"  "Anom34"  "SOI"     "EQSOI"  "BEST"      "TNI"
------------------------------------------------------------------------------
```

5.11 Code 11 Plot Element vs. Index by State

```
------------------------------------------------------------------------------------
Table 19.  Sample output of monthly data summarized by state/territory with corresponding
monthly ENSO indices.  Precipitation, US and Canada.

      St        loc      elem        year    month     pre82       VALm      VALm_US
 1:   BC        CA       PRCP        1960      1         Y         40321     158.74409
 2:   BC        CA       PRCP        1960      2         Y         32910     129.56693
 3:   BC        CA       PRCP        1960      3         Y         30385     119.62598
 4:   BC        CA       PRCP        1960      4         Y         27480     108.18898
 5:   BC        CA       PRCP        1960      5         Y         26997     106.2874
 6:   BC        CA       PRCP        1960      6         Y         11381     44.80709
```

```
      sumVALsqd_US    VALsqm_US     zerrec   recs    zblank  zerobs     obs   daysinmo
 1:   192.12355       1613.4087      429     863       0      429       863      868
 2:   142.06836       1054.816       467     806       0      467       806      812
 3:   86.56722        882.4508       452     849       0      452       849      868
 4:   79.6731         826.5286       488     856       0      488       856      870
 5:   61.02212        601.5856       404     875       0      404       875      899
 6:   15.57885        109.3986       570     838       0      570       838      840

      misinm   clsdinm   stns    avgVALm_US      Nino12     Anom12    Nino3     Anom3
 1:      5        0       28     0.18394449       24.23      -0.31     25.31    -0.35
 2:      6        0       28     0.16075301       25.68      -0.34     25.93    -0.47
 3:     19        0       28     0.14090222       26.24      -0.18     26.87    -0.29
 4:      8        6       29     0.12638899       24.43      -1.04     27.15    -0.33
 5:     24        0       29     0.12147132       23.33      -0.94     26.71    -0.45
 6:      2        0       28     0.05346908       21.71      -1.3      25.86    -0.64

      Nino4       Anom4     Nino34      Anom34     SOI      EQSOI       BEST      TNI
 1:    27.62       -0.62      26.27       -0.29    -1.5       -1.1       1.51    -0.945
 2:    27.44       -0.65      26.29       -0.45    -1.2       -0.7       0.74    -0.668
 3:    27.75       -0.45      26.98       -0.25     0.4        0.4      -0.07    -1.399
 4:    28.01       -0.44      27.49       -0.22    -0.2       -0.1       0.7     -1.911
 5:    28.42       -0.33      27.68       -0.13     0.4        0.7      -0.39    -0.373
 6:    28.33       -0.46      27.24       -0.35     2.9        0.3      -0.75    -1.149
----------------------------------------------------------------------------------------------
```

Code 11 summarizes by state, greatly reducing the data size. This code is intended to be highly customizable. The example given is a simplified illustration. Yearly data is used to select stations with records in all 57 base and recent years, which adds consistency to the location of observations across time so that comparisons of yearly results are meaningful. The detailed station inventories could be used to make this type of selection, but reading yearly data has the advantage including counts of observations by which to further refine selections. Alternatively, the year-month summary could be used to set a minimum level of completeness for selected months based on monthly observation counts. Many other criteria can be introduced. As the weather element data will be summarized from a station level to a state level, it is important to consider the stations represented by the selections. Strict selections may result in overly sparse records by state or sparseness in relevant regions.

The data to read in will be monthly, bimonthly or trimonthly, depending on the index selection. The MEI is bimonthly and ONI is trimonthly. The vector "IndexName" is defined by the column headings for index values, the same as in Code 10. A loop is coded so that all of the indices can be plotted at once; or the loop can be commented out. If errors are encountered, remember to reset the column headers to the original headers.

The code identifies years prior to 1982 versus years from 1982 on, and base climate period years versus subsequent years. The data is summed preserving the 1982 split but could be modified to retain the base period, or another split specified by a code modification. Outliers labeled to identify the year of each point, can be revised at the left and right limits, according to the overall spread of the plot. Some graph labels are also automated and may need to be refined.

For the example, no adjustment is made for the change in proportion of blank records assumed zero ('zblank') which has been drastic especially since 1982. The plot is color-coded to show points before and after 1982. Erroneous zero entries will be overstated prior to 1982, so that adjustments for unidentified blanks assumed zero would shift points upward. The adjustment should affect the final relationship displayed between the weather element and the ENSO index. The effort to improve completion of records since 1982 will also cause zero records to increase since 1982, over which time zero entries have been somewhat understated.

5.12 Code 12 Map of ENSO Index Regions

This code creates a map of the ENSO Index regions. Details on mapping are covered in Part I.

5.13 Code 13 Costliest Storms

The data on costliest storms was copied from Wikipedia in January 2018. The table below can be copied into excel and saved as a *.csv file 'CostlyStorms.csv' in the base directory as input to this code. The Wikipedia data is updated regularly, and if the input file is updated then the formatting of the bar graph will require updates to the code. Some storm dates are provide which can be used for comparison against ENSO indices.

Table 20. List of Costliest Atlantic Hurricanes. Storms exceeding U.S. $1 Billion, in descending order. Storms that broke the historical record for damages, at the time of the storm's dissipation, are highlighted, showing that the costliest storms have move up the list in large strides that may appear uncharacteristic of inflation or randomness by damages on an unadjusted actual cost level. This table is intended for input to R. [Source : Wikipedia]

	Peak Classification	**Unadjusted**		**(> $1B)**		
Storm	**Hurricane Category**	**Damages**		**Storm #**	**Begin**	**End**
Name	**(0 = Tropical Storm)**	**in U.S. $Billions**	**Year**	**of Year**	**Date**	**Date**
Katrina	**5**	**125**	**2005**	**3**	**823**	**831**
Harvey	**4**	**125**	**2017**	**1**	**817**	**903**
Maria	5	92	2017	3	916	1003
Sandy	3	68.7	2012	2	1022	1102
Irma	5	64.2	2017	2	830	916
Ike	4	38	2008	3	901	915
Wilma	5	27.4	2005	6	1016	1027
Andrew	**5**	**27.3**	**1992**	**1**	**816**	**828**
Ivan	5	26.1	2004	3	902	924
Rita	5	18.5	2005	4	918	926
Charley	4	16.9	2004	1	809	815
Matthew	5	15.1	2016	1	928	1010
Irene	3	14.2	2011	1	821	830
Frances	4	9.8	2004	2	824	910
Hugo	**5**	**9.47**	**1989**	**1**	**910**	**925**
Georges	4	9.37	1998	2	915	1001
Allison	0	8.5	2001	1	604	620
Gustav	4	8.31	2008	2	825	907
Jeanne	3	7.94	2004	4	913	929
Floyd	4	6.5	1999	1	907	919
Mitch	5	6.08	1998	3	1022	1109

Isabel	5	5.5	2003	1	906	920
Fran	3	5	1996	1	823	910
Opal	4	4.7	1995	3	927	1006
Stan	1	3.96	2005	5	1001	1005
Karl	3	3.9	2010	2	914	918
Dennis	4	3.71	2005	1	704	718
Alicia	**3**	**3**	**1983**	**1**	**815**	**821**
Gilbert	5	2.98	1988	1	908	929
Luis	4	2.97	1995	1	828	912
Lee	0	2.8	2011	2	902	907
Isaac	1	2.8	2012	1	821	903
Michelle	4	2.35	2001	2	1029	1106
Agnes	**1**	**2.1**	**1972**	**1**	**614**	**623**
Marilyn	3	2.1	1995	2	912	930
Dean	5	1.95	2007	1		
Alex	2	1.89	2010	1		
Joan	4	1.87	1988	2		
Fifi	2	1.8	1974	1		
Frederic	4	1.71	1979	2		
Dolly	2	1.6	2008	1		
Allen	5	1.57	1980	1		
David	5	1.54	1979	1		
Bob	3	1.51	1991	1		
Juan	1	1.5	1985	2		
Roxanne	3	1.5	1995	4		
Ingrid	1	1.5	2013	1		
Betsy	**4**	**1.43**	**1965**	**1**		
Camille	5	1.42	1969	1		
Elena	3	1.3	1985	1		
Isidore	3	1.28	2002	1		
Lili	4	1.16	2002	2		
Alberto	0	1.03	1994	1		
Emily	5	1.01	2005	2		
Beulah	5	1	1967	1		
Bonnie	3	1	1998	1		

Code

```
# =================================================== #
# ================= CODE CONTENTS =================== #
# =================================================== #

# 5.0                 Set up in R
# 5.1   Code 1        Weather Daily - Loop to Unzip Year by Year
# 5.2   Code 2        Initial Detailed Station Inventories
# 5.3   Code 3        Initial Year-Month Summaries
# 5.4   Code 4        Complete Yearly and Year-Month Summaries
#                            Merged with Station Locations and Inventories
# 5.5   Code 5        Missing Records by Year
# 5.6   Code 6        Multiple Month Indices (MEI and ONI)
# 5.7   Code 7        Weekly Nino Indices
# 5.8   Code 8        State Summaries / Plot Selected Stations
#                         Visual Analysis with Choropleth Maps
#                            5.8.1 Package 'maps' - 48 mainland states
#                            5.8.2 Packages 'ggplot2' and 'fiftystater' - AK & HI insets
# 5.9   Code 9        Combine Monthly Indices
# 5.10  Code 10       Plot Index Time Series
# 5.11  Code 11       Plot Element vs. Index by State
# 5.12  Code 12       Map of ENSO Index Regions
# 5.13  Code 13       Costliest Storms


# Station/Element level:  1 (daily), 2, 3 (yr-mo), 4 (yr-mo & yr), 6 (2 mo & 3 mo), 7 (weekly)
# Country/Element level:  5
# State/Element level:    8, 10
#
# 1. Loop to unzip daily meteorological data year by year.  Long run time.
# 2. Open daily files to detail station inventories, dates open/closed, missing records.
# 3. Open daily files to summarize by year-month.
# 4. Merge station detail into year-month summary; summarize by year
# 5. Yearly summary of total and average counts of missing records, blanks assumed zero,
#      observations, etc.
# 6. Summarize monthly data into two- and three-month periods for comparison to MEI and ONI.
# 7. Open daily files to summarize by seven day periods matching weekly Nino indices.
# 8. Customizable station selections, summarize by state, plot stations and choropleth
# 9. Convert wide formats to long; combine monthly ENSO indices into one .csv file for
#      convenience.
#
# ENSO Indices - bimonthly (MEI), trimonthly (ONI), monthly (Nino, SOI, EQSOI, TNI, BEST),
# and weekly (Nino)

# Daily GHCNDex data - download files by year into a folder specified as the working directory
# ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/
# ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/

# =================================================== #
# ================= SET UP IN R ===================== #
# =================================================== #


########################################
#             BEGIN SETUP              #
########################################

# Set Default Working Directory (Optional)
```

```
setwd("C:/…/Weather")
getwd()

# Remove list to free memory
rm(list=ls())
ls()

# Set directories for downloaded zipped files, base input files, and written output
dirzip <- "C:/…/WeatherZip"
dirbase <- "C:/…/WeatherBase"
diroutput <- "C:/…/WeatherData"

# Load packages data.table, tidyverse, lubridate
library(data.table)   # data.table functions run faster than base R code.
# rbindlist() combines years of weather dataframes in list;
# setnames() updates column headers
library(tidyverse)    # a set of packages for organizing data; package 'readr' to unzip.
library(lubridate)    # days_in_month() gives expected number of records


#................. FUNCTION REPEAT ROWS ...............
rep.row <- function(x,n){
  matrix(rep(x,each=n),nrow=n)
}
#......................................................


Elements <- c('PRCP', 'SNOW', 'SNWD', 'TMAX', 'TMIN', 'WIND')

#----------------- FIVE CORE ELEMENTS ----------------------------------------
#
# SelElem      Element           Unit of Measure        Converted (US)
#
# PRCP Precipitation         tenths of mm           inches
# SNOW Snowfall                  mm                     inches
# SNWD Snow depth              mm                     inches
# TMAX Maximum temperature    tenths of degrees C    degrees Fahrenheit
# TMIN Minimum temperature    tenths of degrees C    degrees Fahrenheit
#
#------------------------------------------------------------------------------
#
# WIND    elements are coded to include:
#
# AWND    Average daily wind speed         (tenths of meters per second)
# WSF1    Fastest 1-minute wind speed      (tenths of meters per second)
# WSF2    Fastest 2-minute wind speed      (tenths of meters per second)
# WSF5    Fastest 5-second wind speed      (tenths of meters per second)
# WSFG    Peak gust wind speed             (tenths of meters per second)
# WSFI    Highest instantaneous wind speed (tenths of meters per second)
# WSFM    Fastest mile wind speed          (tenths of meters per second)
#
#-------------------------------------------------------------------------------


########################################
#             END SETUP                #
########################################

#
==============================================================================================
#
```

```
# ===== CODE 1 ===== LOOP TO UNZIP DAILY METEOROLOGICAL FILES YEAR BY YEAR
==================== #
#
==============================================================================================
#

# Elements                    # View list of elements
# Elements <- Elements[1:2]   # Select subset of elements (option)

# Set directory to file location of downloaded zipped files
setwd(dirzip)
# Put files in directory that will be unzipped and read
gzfiles <- dir(pattern = "*.csv.gz") # creates the list of all the csv files in the directory
gzfiles     # view files selected
# gzfiles <- gzfiles[1:3]  # Select subset of files from the list (option)

dataset <- list() # creates a list that will hold the meteorological data files

########################################
#          BEGIN OUTER LOOP            #
########################################

# OUTER LOOP : Selected years of zipped .GZ daily weather station data files

for (k in 1:length(gzfiles)){
  setwd(dirzip)
  dataset[[k]] <- read_csv(gzfiles[k], col_names = FALSE)
  # Add column names
  colnames(dataset[[k]]) <- c("StationID","date", "elem", "VAL", "MFlag",
                              "QFlag", "SFlag", "Time" )

  # Create data table to save processing time - Subset elements from this table
  dtbl <- as.data.table(dataset[[k]])

  # reduce the large data frame in the list to save memory in the loop
  dataset[[k]] <- 0

  # Create location field (country etc.) to be used to subset data
  dtbl[, loc := substring(StationID, 1, 2)]
  class(dtbl$elem)

  ########################################
  #          BEGIN INNER LOOP            #
  ########################################

  # Inner Loop : all Elements for the unzipped year

  for (L in 1: length(Elements)){
    # 'WIND' will subset several wind elements; otherwise use SelElem
    ifelse(Elements[L] != 'WIND',  SelElem <- Elements[L],
           SelElem <- c("AWND", "WSF1", "WSF2", "WSF5", "WSFG", "WSFI", "WSFM"))

    # Subset US and Canadian data for selected element, so datasets are small enough to write
    subdat <- dtbl[elem %in% SelElem & loc %in% c('US', 'CA')]
    subdat[is.na(subdat)]<- "-"
    # Replace line above to include US Territories
    # subdat <- dtbl[elem %in% SelElem & loc %in% c('US', 'CA', 'AQ', 'CQ', 'GQ',
    #'JQ', 'LQ', 'RQ', 'VQ', 'WQ')]

# Add year and month fields
subdat[, year := as.integer(substring(date, 1, 4))]
subdat[, month := as.integer(substring(date, 5, 6))]
subdat[, monthday := as.character(substring(date, 5, 8))]
# N.B. monthday 0122 appears in csv as 122. 1202 appears in csv as 1202.

# Convert unit of measure specific to selected element.
if(SelElem == 'PRCP'){
  subdat[,VAL_US := (VAL/254)]
}
if(SelElem == 'SNOW' | SelElem == 'SNWD'){
  subdat[,VAL_US := (VAL/25.4)]
}
if(SelElem == 'TMAX' | SelElem == 'TMIN'){
  subdat[,VAL_US := (VAL*0.18) +32]
}
if(Elements[L] == 'WIND'){
  subdat[,VAL_US := (VAL/10)*2.23694]
}

# Sort the files by location and StationID
subdat <- subdat[order(subdat$elem, subdat$StationID), ]

# Create file name according to year of data
yrchar = as.character(subdat[2, 10])

# Name file where daily data will be written out to
filenmday = paste0("USCANday", Elements[L], yrchar, ".csv")
# Write DAILY subsets of data to csv files
setwd(diroutput)
write_csv(subdat, filenmday, col_names=TRUE)

# Remove datasets and unused values to save memory
rm(filenmday, yrchar, subdat)

gc()  # call for garbage can saves memory
  }

  ########################################
  #      END INNER LOOP (Elements)       #
  ########################################

  rm(dtbl)

}

########################################
#        END OUTER LOOP (Years)        #
########################################

rm(dataset, gzfiles, k, L, SelElem)

# END PROGRAM CODE

#

# ========================================================================================== #
# ===== CODE 2 ================ INITIAL STATION INVENTORY ================================== #
# ========================================================================================== #

# Go back to repeat SETUP at top if R has been closed.

# Select subset of elements (option)
```

```
Elements <- c('PRCP', 'SNOW', 'SNWD', 'TMAX', 'TMIN', 'WIND')
# Elements <- Elements[1:2]

# Select base years (eg. 1961 - 1990) for climatology, typically 30 past years
BegBsYr <- 1961
EndBsYr <- 1990

setwd(dirbase)

stnlist <- read_csv('ghcnd-stations.csv', col_names = FALSE)
colnames(stnlist) <- c("StationID","lat", "lon", "elev", "St", "Name", "GSNFlag", "zip" )
stnlist$loc <- as.character(substring(stnlist$StationID, 1,2))
stnlist <- as.data.table(stnlist)
stnsub <- stnlist[, c('StationID', 'loc', 'St', 'lat', 'lon', 'elev')]  #(*Change columns*)
USCANstn <- subset(stnsub, stnsub$loc %in% c('US', 'CA'))
head(USCANstn)


stninv <- read_csv('ghcnd-inventory.csv', col_names = FALSE)
colnames(stninv) <- c("StationID","lat", "lon", "elem", "Open", "Close")
stninv$loc <- as.character(substring(stninv$StationID, 1,2))
stninv <-as.data.table(stninv)
#(*Change columns*)
stninv0 <- stninv[, c('StationID', 'loc','lat', 'lon', 'elem', 'Open', 'Close')]
USCANinv0 <- stninv0[loc %in% c("US", "CA")]
rm(stnlist, stnsub, stninv, stninv0)

# Set working directory to access output of daily csv files
setwd(diroutput)
########################################
#        BEGIN OUTER LOOP            #
########################################
for (w in 1:length(Elements)){
  SelElem <- Elements[w]
  # Select station inventory by element, merge with station list
  USCANinv <- USCANinv0[elem == SelElem]
  stnloc <- as.data.table(full_join(USCANstn, USCANinv[, c('StationID', 'Open', 'Close')],
                                    by = 'StationID'))

  # creates the list of all the csv files in the directory
  csvfiles <- dir(pattern = paste0("USCANday",  SelElem, "*"))
  csvfiles

  # Define list for loop
  daily <- list()

  ########################################
  #        BEGIN INNER LOOP            #
  ########################################

  ##### LOOP: for selected element, loop through all years of daily records

  for (q in 1:length(csvfiles)){

    daily[[q]] <- read_csv(csvfiles[[q]], col_names = TRUE)
    subdat <- as.data.table(daily[[q]])
    daily[[q]] <- 0

    # Summarize by Station the minimum and maximum operation dates
    stnmindt <- subdat[, lapply(.SD, min, na.rm=TRUE), .SDcols='date',
                       by=list(StationID, loc, elem, year)]
    stnmaxdt <- subdat[, lapply(.SD, max, na.rm=TRUE), .SDcols='date',
                       by=list(StationID, loc, elem, year)]
    setnames(stnmindt, 'date', 'mindate')
    setnames(stnmaxdt, 'date', 'maxdate')
    stndates <- as.data.table(full_join(stnmindt, stnmaxdt,
                                        by = c('StationID', 'loc', 'elem', 'year')))

    rm(subdat, stnmindt, stnmaxdt)

    # Collect years into one data frame
    if(q==1){
     stndtall <- stndates
    }
    if(q>1){
      stndtall <- rbind(stndtall, stndates)
    }

    # Remove files to save memory
    rm(stndates)

    # Call garbage can gc() to save memory
    gc()
  }

  ########################################
  #       END INNER LOOP (Years)         #
  ########################################

  rm(csvfiles, daily, w, q)

  # Continue through code to end

  # Sort the records by location, element (for WIND), year and StationID
  stndtall <- stndtall[order(loc, elem, StationID, year),]

  # Summarize minimum and maximum station operation dates for all combined years
  stnminall <- stndtall[, lapply(.SD, min, na.rm=TRUE), .SDcols='mindate',
                        by=list(StationID, loc, elem)]
  stnmaxall <- stndtall[, lapply(.SD, max, na.rm=TRUE), .SDcols='maxdate',
                        by=list(StationID, loc, elem)]
  stndtsum <- as.data.table(full_join(stnminall, stnmaxall,
                        by = c('StationID', 'loc', 'elem')))

  rm(stnminall, stnmaxall)

  # -------- create additional date fields -----------
  stndtsum[,minyear := as.integer(substring(mindate,1,4))]
  stndtsum[,minmo := as.integer(substring(mindate,5,6))]
  stndtsum[,maxyear := as.integer(substring(maxdate,1,4))]
  stndtsum[,maxmo := as.integer(substring(maxdate,5,6))]

  stndtinv <- as.data.table(right_join(stnloc, stndtsum,
                                       by = c('StationID', 'loc')))

  # --------- partial month adjustments ----------------
  stndtinv[, daysmaxmo := days_in_month(as.Date(paste(maxyear, maxmo, 15, sep ="-")))]
  stndtinv[, begminmo := as.integer(paste0(minyear, ifelse(minmo < 10, "0", ""),
                                           minmo, '01'))]
  stndtinv[, endmaxmo := as.integer(paste0(maxyear, ifelse(maxmo < 10, "0", ""),
                                           maxmo, daysmaxmo))]
  stndtinv[, clsdmbeg := (mindate - begminmo)]
```

```
stndtinv[, clsdmend := (endmaxmo - maxdate)]
# eliminate calculation fields
stndtinv[ ,':='(daysmaxmo = NULL, begminmo = NULL, endmaxmo = NULL)]
# Compare minimum date with ghcnd-inventory station open year
stndtinv[, clsdmbeg := ifelse(Open < minyear, 0, clsdmbeg)]
stndtinv[, clsdmend := ifelse(Close > maxyear, 0, clsdmend)]
stndtinv[, clsdinm := (clsdmbeg + clsdmend)]

# ---------- full month adjustments ----------------
# ....... Set up count of days in full months before and after station in operation .....
# start with days in months for 'regular' year (non leap year) - then add leap year adj
dys <- days_in_month(as.Date(paste(1961, seq(1:12), 15, sep ="-")))
dysleap <- c(dys, 1)
repdys <- rep.row(dysleap, nrow(stndtinv))
repmos <- rep.row(seq(1:12), nrow(stndtinv))
# ......................................................................................
# Count mins
output <- matrix(0, nrow(stndtinv), 13)
for(i in 1:nrow(stndtinv)){
  output[i,1:12] <- (repmos[i,1:12] < stndtinv$minmo[i])
}
# Adjust days in February for leap years: output[,2] is 0 or 1 for second month February
output[,13] <- leap_year(stndtinv$minyear)*(output[,2])
daysout <- repdys*output
stndtinv$clsdbef <- apply(daysout, 1, sum)
# Count maxs
output <- matrix(0, nrow(stndtinv), 13)
for(i in 1:nrow(stndtinv)){
  output[i,1:12] <- (repmos[i,1:12] > stndtinv$maxmo[i])
}
# Adjust days in February for leap years: output[,2] is 0 or 1 for second month February
output[,13] <- leap_year(stndtinv$maxyear)*(output[,2])
daysout <- repdys*output
stndtinv$clsdaft <- apply(daysout, 1, sum)
# Compare minimum date with ghcnd-inventory station open year
stndtinv[, clsdbef := ifelse(Open < minyear, 0, clsdbef)]
stndtinv[, clsdaft := ifelse(Close > maxyear, 0, clsdaft)]
stndtinv[, clsdfulm := (clsdbef + clsdaft)]

# ------- count base and recent years -------------
stndtall[, bsyrct:= ifelse(year >= BegBsYr & year <= EndBsYr, 1, 0)]
stndtall[, rcyrct:= ifelse(year > EndBsYr, 1, 0)]
stnyrct <- stndtall[, lapply(.SD, sum, na.rm=TRUE), .SDcols=c('bsyrct', 'rcyrct'),
                    by=list(StationID, loc, elem)]
stnyrct[ ,bryrct := (bsyrct + rcyrct)]
stndtfin <- as.data.table(full_join(stndtinv, stnyrct,
                     by = c('StationID', 'loc', 'elem')))

rm(stndtall, stndtsum, stndtinv, stnyrct)
rm(dys, dysleap, daysout, repdys, repmos, output, i)

# Sort the records by location, element (for WIND) and StationID
stndtfin <- stndtfin[order(loc, elem, StationID),]


# Fill in missing State (St) from StationID - valid for recent years ID naming convention
stndtfin$St <- ifelse(is.na(stndtfin$St), as.character(substring(stndtfin$StationID, 4, 5)),
                      as.character(stndtfin$St))
# If State (St) filled in from StationID, still missing lat, lon, and elev (1998-2016)

attach(stndtfin)
```

```
  stndtfin$begbs <- ifelse(minyear<=BegBsYr & maxyear>=BegBsYr, BegBsYr,
                           ifelse(maxyear<BegBsYr, 0, ifelse(minyear>EndBsYr, 0, minyear)))
  stndtfin$endbs <- ifelse(minyear<=EndBsYr & maxyear>=EndBsYr, EndBsYr,
                           ifelse(minyear>EndBsYr, 0, ifelse(maxyear<BegBsYr, 0, maxyear)))
  stndtfin$begrc <- ifelse(maxyear<=EndBsYr, 0, ifelse(minyear <=EndBsYr+1,
                                                       EndBsYr+1, minyear))
  stndtfin$endrc <- ifelse(maxyear<=EndBsYr, 0, maxyear)
  attach(stndtfin)
  stndtfin$bsspan <- ifelse((begbs == 0 | endbs == 0), 0, endbs - begbs + 1)
  stndtfin$rcspan <- ifelse((begrc == 0 | endrc == 0), 0, endrc - begrc + 1)
  stndtfin$brspan <- stndtfin$bsspan + stndtfin$rcspan
  # stndtfin$spanyrs <- maxyear - minyear + 1
  # eliminate calculation fields
  stndtfin[ ,':='(begbs = NULL, endbs = NULL, begrc = NULL, endrc = NULL)]
  # Name files to write station inventories
  filenmstns <- paste0("_USCANstndt", SelElem, "df.csv")
  # Write station inventories to csv files
  setwd(diroutput)
  write_csv(stndtfin, filenmstns, col_names=TRUE)

  rm(stndtsum, stndtfin, stndtinv, stndtall)
  rm(filenmstns, csvfiles)

  gc()  # Call garbage can to spare memory

}
########################################
#     END OUTER LOOP (Elements)        #
########################################
rm(USCANstn, USCANinv0, USCANinv, stninv, stnloc)
# rm(BegBsYr, EndBsYr)


##### END Program Code


# ========================================================================================== #
# ===== CODE 3 ========= INITIAL YEAR MONTH SUMMARY (from Daily files) ===================== #
# ========================================================================================== #

# Go back to repeat SETUP at top if R has been closed.

# # Set working directory to access output of daily csv files
setwd(diroutput)

Elements <- c('PRCP', 'SNOW', 'SNWD', 'TMAX', 'TMIN', 'WIND')

########################################
#         BEGIN OUTER LOOP             #
########################################

# OUTER LOOP: Loop through list of selected weather elements

for (z in 1:length(Elements)){

  SelElem <- Elements[z]
  # creates the list of all the csv files in the directory
  csvfiles <- dir(pattern = paste0("USCANday",  SelElem, "*"))
  csvfiles   # View list of selected file names
  #csvfiles <- csvfiles[58]  # Select years
```

```
# Define list for inner loop
daily <- list()
########################################
#        BEGIN INNER LOOP              #
########################################

# INNER LOOP: for a given element, loop through all years of daily records

for (q in 1:length(csvfiles)){

  daily[[q]] <- read_csv(csvfiles[[q]], col_names = TRUE)
  subdat <- as.data.table(daily[[q]])
  daily[[q]] <- 0

  # Create additional field columns
  subdat[, VALsqd_US := (VAL_US)^2]
  subdat[, zerrec := (VAL <= 0) + 0]
  subdat[, recs := 1]
  subdat[, zblank := ifelse(MFlag == 'P', 1, 0)]
  subdat[, zerobs := (zerrec - zblank)]
  subdat[, obs := (recs - zblank)]

  # Select columns to summarize based on SelElem (*Change columns*)
  yrmocol <- c('VAL', 'VAL_US', 'VALsqd_US', 'zerrec', 'recs', 'zblank',   'zerobs', 'obs')


  # Aggregate data into monthly summaries
  yrmosum <- subdat[, lapply(.SD, sum, na.rm=TRUE), .SDcols=yrmocol,
                  by=list(StationID, loc, elem, year, month)]
  setnames(yrmosum, "VAL", "VALm")
  setnames(yrmosum, "VAL_US", "VALm_US")
  setnames(yrmosum, "VALsqd_US", "sumVALsqd_US")
  # Add fields and reorganize columns
  firstcols <- yrmosum[, StationID:sumVALsqd_US]  # reorganize columns
  firstcols[,VALsqm_US := (VALm_US)^2]
  lastcols <- yrmosum[,zerrec:obs]                # reorganize columns
  yrmosum <- cbind(firstcols, lastcols)
  yrmosum[,daysinmo := days_in_month(as.Date(paste(yrmosum$year,
                                         yrmosum$month, 15, sep ="-")))]
  yrmosum[,misclsd := (daysinmo - obs)]
  rm(firstcols, lastcols)

  #yrmosum[,MEANd_US:= VALm_US/obs]
  rm(subdat)

  rm(yrmocol)

  # Collect years into one data frame
  if(q==1){
    yrmoall <- yrmosum
  }
  if(q>1){
    yrmoall <- rbind(yrmoall, yrmosum)
  }

  # Remove files to save memory
  rm(yrmosum)

  # Call garbage can gc() to save memory
  gc()
  }
  ########################################
  #       END INNER LOOP (Years)         #
  ########################################

  # Sort the files by location and StationID
  yrmoall <- yrmoall[order(elem, year, StationID), ]

  # Name files to write all years of monthly summarized data out to
  filenmyrmo <- paste0("_USCANyrmo0", SelElem, "df.csv")
  # Write MONTHLY subsets of data to csv files
  write_csv(yrmoall, filenmyrmo, col_names=TRUE)

  rm(yrmoall)
  rm(filenmyrmo)
  rm(csvfiles)

}

########################################
#     END OUTER LOOP (Elements)        #
########################################

# Clear variables to move on to next code
rm(daily, q, z, SelElem)

#### End program code


# ========================================================================================= #
# ===== CODE 4 ====== MERGE STATION INVENTORIES/LOCATIONS TO YEAR MONTH SUMMARY =========== #
# ========================================================================================= #
# ================================== SUMMARIZE BY YEAR ==================================== #
# ========================================================================================= #

# Go back to repeat SETUP at top if R has been closed.

# Set working directory to access output of initial year-month summary and station inventory
setwd(diroutput)

Elements <- c('PRCP', 'SNOW', 'SNWD', 'TMAX', 'TMIN', 'WIND')
# Elements <- Elements[1:2] # Select elements (option)

########################################
#         BEGIN SINGLE LOOP            #
########################################

# Loop through list of selected weather elements

for (j in 1:length(Elements))
{
  SelElem <- Elements[j]

  # Find file name to read in - year month data summary
  fileyrmo <- paste0("_USCANyrmo0", SelElem, "*")
  yrmofiles <- dir(pattern = fileyrmo) # creates the list of the files in the directory
  yrmofiles   # view files selected

  # Find file name to read in - station dates of operation and inventories
  filenmstn <- paste0("_USCANstndt", SelElem, "*")
  stnfiles <- dir(pattern = filenmstn) # creates the list of the files in the directory
```

```
 stnfiles   # view files selected

 # Read in Data Files - yearly station data, station dates of operation, station locations
 yrmodat <- read_csv(yrmofiles[1], col_names = TRUE)
 stndates <- read_csv(stnfiles[1], col_names = TRUE)

 # Remove variables to clean up environment
 rm(fileyrmo, filenmstn, yrmofiles, stnfiles)

 # Create data tables to process faster
 yrmodat <- as.data.table(yrmodat)
 stndates <- as.data.table(stndates)

 # ---------- merge partial month adj to 'Year Month' data ----------------
 stnmins <- stndates[, .(StationID, elem, minyear, minmo, clsdmbeg)]  #(*Change columns*)
 stnmins <- stnmins[clsdmbeg > 0]
 setnames(stnmins, 'minyear', 'year')
 setnames(stnmins, 'minmo', 'month')

 stnmaxs <- stndates[, .(StationID, elem, maxyear, maxmo, clsdmend)] #(*Change columns*)
 stnmaxs <- stnmaxs[clsdmend > 0]
 setnames(stnmaxs, 'maxyear', 'year')
 setnames(stnmaxs, 'maxmo', 'month')

 yrmomin <- as.data.table(left_join(yrmodat, stnmins,
                          by = c('StationID', 'elem', 'year', 'month')))
 yrmofin <- as.data.table(left_join(yrmomin, stnmaxs,
                          by = c('StationID', 'elem', 'year', 'month')))
 yrmofin[is.na(yrmofin)]<- 0
 rm(stnmins, stnmaxs, yrmomin)
 rm(yrmodat)

 # _______ MERGE STATES / PROVINCES / COORDINATES / STATION MIN/MAX DATES __________________

 # Add state, coordinates, range of station operation dates (min and max)
 yrmodet <-  as.data.table(right_join(stndates[, StationID:maxdate], yrmofin,
                           by = c('StationID', 'loc', 'elem')))  #(*Change Columns*)
 # keep yrmofin for yearly summary
 # yrmosumdet <- yrmodet[,c(1:2, 21:26, 3:20)]  #(*Change Columns*)
# yrmosumdet <- cbind(yrmodet[,StationID:loc], yrmodet[,St:maxdate],
#                    yrmodet[,elem:clsdmend]) #misinm]) #(*Change Columns*)
 yrmodet$clsdmend <- as.integer(yrmodet$clsdmend)
 yrmodet$clsdmbeg <- as.integer(yrmodet$clsdmbeg)
 yrmodet[, misinm := misclsd - (clsdmbeg + clsdmend)]
 yrmodet[, clsdinm := (clsdmbeg + clsdmend)]
 # Remove columns
 yrmodet[ ,':=' (clsdmbeg = NULL, clsdmend = NULL, misclsd = NULL)]

 # Name file to write data out to
 fileyrmoinv <- paste0("_USCANyrmoinv", SelElem, "df.csv")
 # Write YEAR MONTH inventories of data to csv files
 write_csv(yrmodet, fileyrmoinv, col_names=TRUE)
 rm(yrmodet, fileyrmoinv)

 # Continue to sum by year, using year-month data prior to station location merge
 # Add count field for months
 yrmofin[, count := 1]
 yrmofin$clsdmbeg <- as.integer(yrmofin$clsdmbeg)
 yrmofin$clsdmend <- as.integer(yrmofin$clsdmend)
 # Select data columns to summarize yearly based on SelElem. Eliminate month field
 yrcol <- names(yrmofin)[c(6:19)]   #(*Change columns)
```

```
yrsum <- yrmofin[, lapply(.SD, sum, na.rm=TRUE), .SDcols=yrcol,
                 by=list(StationID, loc, elem, year)]
setnames(yrsum, "VALm", "VALy")
setnames(yrsum, "VALm_US", "VALy_US")
setnames(yrsum, "VALsqm_US", "sumVALsqm_US")
setnames(yrsum, "daysinmo", "daysum")
setnames(yrsum, 'count', 'monthct')

# Rearrange column order and insert new field columns
firstcols <- yrsum[,StationID:sumVALsqm_US]
firstcols[,VALsqy_US := VALy_US^2]
midcols <- yrsum[,zerrec:daysum]
midcols[,daysinyr := 365+leap_year(yrsum$year)*1]
yrsum <- cbind(firstcols, midcols, yrsum[,misclsd:monthct])
# Sort data
yrsum <- yrsum[order(StationID, elem, year),]
# Remove data sets to free space
rm(firstcols, midcols, yrcol)
# Keep yrmofin to spread monthly observations to a yearly format

# ------ merge full month adjustments to 'Year' data ----------------
stnyrmins <- stndates[, c("StationID", "elem", "minyear", "clsdbef")]    #(*Change Columns*)
stnyrmins <- stnyrmins[clsdbef > 0]
setnames(stnyrmins, 'minyear', 'year')

stnyrmaxs <- stndates[, c("StationID", "elem", "maxyear", "clsdaft")]    #(*Change Columns*)
stnyrmaxs <- stnyrmaxs[clsdaft > 0]
setnames(stnyrmaxs, 'maxyear', 'year')

yrmin <- as.data.table(left_join(yrsum, stnyrmins,
                        by = c('StationID', 'elem', 'year')))
yrfin <- as.data.table(left_join(yrmin, stnyrmaxs,
                        by = c('StationID', 'elem', 'year')))
rm(stnyrmins, stnyrmaxs, yrmin)
rm(yrsum)
yrfin[is.na(yrfin)]<- 0

yrfin[, clsdinm := (clsdmbeg + clsdmend)]
yrfin[, clsdfulm := (clsdbef + clsdaft)]
yrfin[, clsdall := clsdinm + clsdfulm]

yrfin[, misinm := (misclsd - clsdinm)]
yrfin[, misfulm := (daysinyr - daysum - clsdfulm)]
yrfin[, misall := (misinm + misfulm)]

# Reduce columns
yrfin[ ,':=' (clsdmbeg = NULL, clsdmend = NULL)]
yrfin[ ,':=' (clsdbef = NULL, clsdaft = NULL)]
yrfin[ , misclsd := NULL]

# Spread monthly observations in a 12 column grid for each station - year
# yrmoobs <- yrmofin[, .(StationID, elem, year, month, obs)]
# yrmogrid <- spread(yrmoobs, month, obs)
# yrmogrid[is.na(yrmogrid)] <- 0
# rm(yrmoobs)
# colnames(yrmogrid) <- c('StationID', 'elem', 'year',  'mo01',  'mo02', 'mo03', 'mo04',
# 'mo05', 'mo06', 'mo07', 'mo08', 'mo09', 'mo10', 'mo11',  'mo12')
# Compare to 'obs' data check
# yrmogrid[, yrobs := mo01+mo02+mo03+mo04+mo05+mo06+mo07+mo08+mo09+mo10+mo11+mo12]
# yrmogrid <- yrmogrid[order(StationID, elem, year),]
```

```
  #yrsumgrid <- as.data.table(left_join(yrfin, yrmogrid, by = c('StationID', 'elem', 'year')))
  #rm(yrfin, yrmogrid)

  stndat <- cbind(stndates[,bsyrct:brspan], stndates[,StationID:maxdate])
  rm(stndates)
  yrdet <- as.data.table(left_join(yrfin, stndat, by = c('StationID', 'loc', 'elem')))
  rm(yrfin, stndat)
  #rm(stndates)

  # Reorder data to organize for output
  #(*Change Columns*)
  yrsumdet <- cbind(yrdet[,StationID:loc], yrdet[,St:maxdate], yrdet[,elem:brspan])
  rm(yrdet)

  # Sort data
  yrsumdet <- yrsumdet[order(elem, year, StationID), ]

  # Name file to write data out to
  fileyrdet <- paste0("_USCANyrinvgrid", SelElem, "df.csv")
  # Write YEARLY summary to csv files
  write_csv(yrsumdet, fileyrdet, col_names=TRUE)
  rm(fileyrdet, yrsumdet)

  gc()

}
########################################
#         END LOOP (Elements)          #
########################################

##### End Program Code


# ============================================================================================ #
# ===== CODE 5 ============== Missing Records Summary by Year (loc) ======================= #
# ============================================================================================ #

# Go back to repeat SETUP at top if R has been closed.

# Set working directory to access initial year-month summary and station inventory
setwd(diroutput)

########################################
#            BEGIN LOOP                #
########################################

# Loop through yearly inventories, sum all stations by year

yrfiles <- dir(pattern = "_USCANyrinvgrid") # creates the list of the files in the directory
yrfiles  # view files selected

for (g in 1:length(yrfiles)){

  # Read in Data Files - yearly station data, station dates of operation, station locations
  yrdat <- read_csv(yrfiles[g], col_names = TRUE)
  yrdat <- as.data.table(yrdat)
  # Reduce columns
  yrdat <- yrdat[, StationID:misall]  # Select consecutive columns

  ###################################################
  # Summarize missing records by year and loc
  yrdat[,stnct := 1]
  miscols <- names(yrdat)[c(18:32)]   #(*Change columns*)
  misrecsum <- yrdat[, lapply(.SD, sum, na.rm=TRUE), .SDcols=miscols,
                     by=list(loc, elem, year)]
  misrecsum <- as.data.table(misrecsum)
  misrecsum <- misrecsum[order(loc, elem, year),]


  if (g==1){
    misrecall <- misrecsum
  }
  if (g > 1){
    misrecall <- rbind(misrecall, misrecsum)
  }

  rm(misrecsum, yrdat)
}

# Define percentage fields (Note zerobs-to-obs has a different denom)
misrecall[, pctobsyr := (obs/daysinyr)]
misrecall[, pctmisyr := (misall/daysinyr)]
misrecall[, pctclsdyr := (clsdall/daysinyr)]
misrecall[, pctzblkyr := (zblank/daysinyr)]
misrecall[, pctzeroyr := (zerobs/daysinyr)]
misrecall[, pctzerobs := (zerobs/obs)]
setnames(misrecall, 'daysum', 'stndays')
setnames(misrecall, 'monthct', 'stnmos')
setnames(misrecall, 'daysinyr', 'stndysinyr')

# Rearrange column order
misrecfin <- cbind(misrecall[, loc:year], misrecall[, pctzerobs:pctobsyr],
                   misrecall[, .(stnct)], misrecall[, zerrec:misall])


#misrecfin[, daysyravg := stndysinyr/stns] # check it is 365 or 366
misrecfin[, obsavg := (obs/stnct)]
misrecfin[, clsdavg := (clsdall/stnct)]
misrecfin[, misavg := (misall/stnct)]
misrecfin[, zblankavg := (zblank/stnct)]
misrecfin[, zeroavg := (zerobs/stnct)]
misrecfin[, mosavg := (stnmos/stnct)]


########################################
#         END LOOP (Elements)          #
########################################

# Write out Missing Record Summary for All Elements
write_csv(misrecall, "_USCANmisrecELEMdf.csv", col_names=TRUE)

# Reopen option
# misrecall <- read_csv("_USCANmisrecELEMdf.csv", col_names=TRUE)
# misrecall <- as.data.table(misrecall)

mistbl <- misrecall[,c('loc', 'elem', 'year', 'pctmisyr')]
clsdtbl <- misrecall[,c('loc', 'elem', 'year', 'pctclsdyr')]
obstbl <- misrecall[,c('loc', 'elem', 'year', 'pctobsyr')]
stntbl <- misrecall[,c('loc', 'elem', 'year', 'stnct')]
zerobstbl <- misrecall[,c('loc', 'elem', 'year', 'pctzerobs')]

mistbl <- mistbl[elem %in% c('PRCP', 'SNOW', 'SNWD', 'TMAX', 'TMIN')]
```

```
mistblw <- spread(mistbl, elem, pctmisyr)
mistblw[is.na(mistblw)] <- ""
mistblw

clsdtbl <- clsdtbl[elem %in% c('PRCP', 'SNOW', 'SNWD', 'TMAX', 'TMIN')]
clsdtblw <- spread(clsdtbl, elem, pctclsdyr)
clsdtblw[is.na(clsdtblw)] <- ""
clsdtblw

obstbl <- obstbl[elem %in% c('PRCP', 'SNOW', 'SNWD', 'TMAX', 'TMIN')]
obstblw <- spread(obstbl, elem, pctobsyr)
obstblw[is.na(obstblw)] <- ""
obstblw

zerobstbl <- zerobstbl[elem %in% c('PRCP', 'SNOW', 'SNWD', 'TMAX', 'TMIN')]
zerobsw <- spread(zerobstbl, elem, pctzerobs)
zerobsw[is.na(zerobsw)] <- ""
zerobsw

write_csv(obstblw, 'pctmisrec.csv', col_names=TRUE)
write_csv(mistblw, 'pctmisrec.csv', col_names=TRUE)
write_csv(clsdtblw, 'pctclsdrec.csv', col_names=TRUE)
write_csv(zerobsw,  'pctzerobs.csv', col_names=TRUE)

##### END PROGRAM CODE



# ========================================================================================== #
# ===== CODE 6 ========== BIMONTHLY (MEI) and TRIMONTHLY (ONI) INDICES ===================== #
# ========================================================================================== #

# Go back to repeat SETUP at top if R has been closed.

Elements <- c('PRCP', 'SNOW', 'SNWD', 'TMAX', 'TMIN', 'WIND')

# Open station level summaries by year month; sum by two- and three- month periods.
# The MultiVariate ENSO Index (MEI) is bimonthly (two months).
# The Oceanic Nino Index (ONI) is trimonthly (three months).

# Read in MEI and ONI from base directory
setwd(dirbase)
MEI <- read_csv('MEI_Index.csv', col_names = TRUE)
colnames(MEI) <- c('year', 'DecJan', 'JanFeb', 'FebMar', 'MarApr', 'AprMay', 'MayJun',
                    'JunJul',  'JulAug',  'AugSep', 'SepOct', 'OctNov', 'NovDec')
ONI <- read_csv('ONI_Index.csv', , col_names = TRUE)
head(ONI)

# use functions from tidyvers package to convert index lists from wide to long
MEIlong <- gather(MEI, "bimo", "MEI", 2:13)
head(MEIlong)
ONIlong <- gather(ONI, "trimo", "ONI", 2:13)
head(ONIlong, 10)
setnames(ONIlong, 'Year', 'year')
rm(MEI, ONI)

# Create labels for bimonthly and trimonthly aggregations
bimolabel <- data.frame(k = seq(1, 12, 1),
                        bimo = c('JanFeb','FebMar', 'MarApr', 'AprMay', 'MayJun', 'JunJul',
                                 'JulAug', 'AugSep', 'SepOct', 'OctNov', 'NovDec', 'DecJan'))
trimolabel <- data.frame(k = seq(1, 12, 1),
                         trimo = c('JFM','FMA', 'MAM', 'AMJ', 'MJJ', 'JJA', 'JAS', 'ASO',
                         'SON', 'OND', 'NDJ', 'DJF'))


# Set working directory to access year-month summaries
setwd(diroutput)
########################################
#           BEGIN MAIN LOOP            #
########################################

for (h in 1:length(Elements))
{
  SelElem <- Elements[h]

  fileyrmo <- paste0("_USCANyrmoinv", SelElem, "df.csv")
  fileyrmo
  filestndt <- paste0("_USCANstndt", SelElem, "df.csv")
  filestndt
  #  yrmofiles <- dir(pattern = fileyrmo) # creates the list of the files in the directory
  #  yrmofiles  # view files selected

  setwd(diroutput)
  # Read in Data files - year month data with missing records and closed dates inventories
  yrmoall <- read_csv(fileyrmo[1], col_names = TRUE)
  # Select columns and place in data table for efficiency
  yrmodat <- as.data.table(yrmoall)   # 8,579,187
  rm(yrmoall, fileyrmo)
  minyr <- min(yrmodat$year)
  maxyr <- max(yrmodat$year)

  yrmodat[, monthct := 1]   # Count months summed in loop
  yrmodat$monthX <- yrmodat$month
  # Create 'Month 13' data as January of next year, 'Month 14' as February of next year
  wrapmo13 <- yrmodat[month == 1]
  wrapmo14 <- yrmodat[month == 2]
  wrapmo13$monthX <- 13
  wrapmo13$year <- wrapmo13$year - 1
  wrapmo14$monthX <- 14
  wrapmo14$year <- wrapmo14$year - 1
  loopsum <- rbind(yrmodat, wrapmo13, wrapmo14)
  loopsum <- loopsum[year >= minyr]
  loopsum <- loopsum[order(elem, StationID, year, monthX), ]
  rm(yrmodat)

  # Create lists for loop
  bimosum <- list()
  trimosum <- list()
  # Select columns of values to sum in loop
  sumcols <- names(loopsum)[c(14:26)]   #(*Change columns*)
  ########################################
  #       BEGIN MINOR LOOP               #
  ########################################
  for (k in 1:12){
    mo1 <- as.character(k)
    mo2 <- as.character(k+1)
    mo3 <- as.character(k+2)
    #mo2lab <- ifelse(k == 12, 1, mo2)
    #mo3lab <- ifelse(k == 12, 2, mo3)
    bimosub <- loopsum [monthX == mo1 | monthX == mo2]
    trimosub <- loopsum[monthX %in% c(mo1, mo2, mo3)]
    bimosum[[k]] <- bimosub[, lapply(.SD, sum, na.rm=TRUE), .SDcols=sumcols,
                             by=list(StationID, loc, elem, year)]
    trimosum[[k]] <- trimosub[, lapply(.SD, sum, na.rm=TRUE), .SDcols=sumcols,
                               by=list(StationID, loc, elem, year)]
    bimosum[[k]]$ord <- k
    trimosum[[k]]$ord <- k
    bimosum[[k]]$bimo = bimolabel[k,2]
    trimosum[[k]]$trimo = trimolabel[k,2]
  }
  ########################################
  #         END MINOR LOOP               #
  ########################################

  # Continue through code to end

  rm(wrapmo13, wrapmo14, mo1, mo2, mo3, k)
  rm(loopsum, bimosub, trimosub, sumcols)

  bimosumall <- rbindlist(bimosum)
  trimosumall <- rbindlist(trimosum)
  setnames(bimosumall, "daysinmo", "daysum")
  setnames(trimosumall, "daysinmo", "daysum")
  setnames(bimosumall, "VALm", "VAL2m")
  setnames(trimosumall, "VALm", "VAL3m")
  setnames(bimosumall, "VALm_US", "VAL2m_US")
  setnames(trimosumall, "VALm_US", "VAL3m_US")
  setnames(bimosumall, "VALsqm_US", "sumVALsqm_US")
  setnames(trimosumall, "VALsqm_US", "sumVALsqm_US")
  rm(bimosum, trimosum)

  class(bimosumall$year)
  bimosumall$delete <- ifelse((bimosumall$year == maxyr & bimosumall$ord == 12), 'Y', 'N')
  trimosumall$delete <- ifelse((trimosumall$year == maxyr & trimosumall$ord == 11), 'Y',
                        ifelse((trimosumall$year == maxyr & trimosumall$ord == 12), 'Y', 'N'))

  bimofin <- subset(bimosumall, bimosumall$delete == 'N')
  trimofin <- subset(trimosumall, trimosumall$delete == 'N')
  bimofin[, delete:=NULL]
  trimofin[, delete:=NULL]
  rm(bimosumall, trimosumall)
  #nrow(bimosumfin)  # 8,750,322 PRCP
  #nrow(trimosumfin) # 8,846,206 PRP
  bimofin$year <- as.integer(bimofin$year)
  bimofin$bimo <- as.character(bimofin$bimo)
  trimofin$year <- as.integer(trimofin$year)
  trimofin$trimo <- as.character(trimofin$trimo)

  # Merge MEI (bimo) and ONI (trimo) indices with summed data
  bimoInd <- as.data.table(left_join(bimofin, MEIlong,
                           by = c('year', 'bimo')))
  trimoInd <- as.data.table(left_join(trimofin, ONIlong,
                            by = c('year', 'trimo')))
  rm(bimofin, trimofin)

  # Merge Station inventories / states / coordinates
  setwd(diroutput)
  stndates <- read_csv(filestndt, col_names = TRUE)
  stndates  <- as.data.table(stndates)

  bimoIndSt <- as.data.table(right_join(stndates[, StationID:maxdate], bimoInd,
                              by = c('StationID', 'loc', 'elem'))) #(*Change columns*)#
  rm(bimoInd)
```

```
  trimoIndSt <- as.data.table(right_join(stndates[, StationID:maxdate], trimoInd,
                              by = c('StationID', 'loc', 'elem'))) #(*Change columns*)#
  rm(trimoInd)
  rm(stndates)

  setwd(diroutput)
  bimofilenm <- paste0('_USCANbimo', SelElem, 'df.csv')
  trimofilenm <- paste0('_USCANtrimo', SelElem, 'df.csv')
  write_csv(bimoIndSt, bimofilenm, col_names = TRUE)
  write_csv(trimoIndSt, trimofilenm, col_names = TRUE)

  rm(bimofilenm, trimofilenm, filestndt)
  rm(bimoIndSt, trimoIndSt)

}

########################################
#       END MAIN LOOP (Elements)       #
########################################

rm(bimolabel, trimolabel, MEIlong, ONIlong)
rm(h, maxyr, minyr)

##### End Program Code



# =========================================================================================== #
# ===== CODE 7 ========================== WEEKLY INDICES ==================================== #
# =========================================================================================== #

# Go back to repeat SETUP at top if R has been closed.

Elements <- c('PRCP', 'SNOW', 'SNWD', 'TMAX', 'TMIN', 'WIND')

# For weekly Nino Indices choose BegDat <- "1989/12/31"
# (Weekly SST data starts week centered on 1990/01/03)
# -------------------------------------------------------
BegDate <- "1961/01/01"    # 1961/01/01 gives same grouping as 1989/12/31
EndDate <- "2017/12/31"
# --------- weekly labels for grouping sums -----------
dte = seq(as.Date(BegDate), as.Date(EndDate), "days")  # sequence of dates from beg to end
nodys <- length(dte)   # Number of days in sequence
noweeks <- as.integer(nodys/7)   # Number of full weeks in sequence
remdys <- nodys%%7             # Remaining days, not full weeks
weeklabel <- data.frame(weekno = c(rep(1:noweeks, each = 7), rep(noweeks+1, remdys)), dte =
seq(as.Date(BegDate), as.Date(EndDate), "days"))
weeklabel$chardate <- paste0(as.character(substring(weeklabel$dte, 1, 4)),
as.character(substring(weeklabel$dte, 6, 7)), as.character(substring(weeklabel$dte, 9,10)))
weeklabel$date <- as.integer(weeklabel$chardate)
startlab <- as.Date(BegDate)+3      # Weeks are labeled by middle day
centerdates <- seq(as.Date(startlab), by = "7 days", length.out = noweeks)  # middle day for
each week of sequence
weeklabel$ctrweek <- c(rep(centerdates, each = 7), rep(as.Date(tail(centerdates,1))+7,
remdys))  # label all days
weeklabel$yrgrp <- substring(weeklabel$ctrweek, 1, 4)
weeklabel <- weeklabel[,c(1,4:6)]
rm(BegDate, EndDate, centerdates, dte, nodys, remdys, startlab, noweeks)
# -------------------------------------------------------

########################################
```

```
#          BEGIN OUTER LOOP            #
########################################

for (z in 1:length(Elements)){

  # Set working directory to access output of daily csv files
  setwd(diroutput)

  SelElem <- Elements[z]
  # creates the list of all the csv files in the directory
  csvfiles <- dir(pattern = paste0("USCANday",  SelElem, "*"))
  csvfiles <- csvfiles[30:58]  # select years since 1989 to match Nino indices
  # csvfiles

  daily <- list()
  ########################################
  #    BEGIN FIRST INNER LOOP            #
  ########################################
  # for selected element, loop through all years of daily records

  for (q in 1:length(csvfiles)){

    # Set working directory to access output of daily csv files
    setwd(diroutput)

    daily[[q]] <- read_csv(csvfiles[[q]], col_names = TRUE)
    subdat <- as.data.table(daily[[q]])
    daily[[q]] <- 0

    # Create additional field columns
    subdat[, VALsqd_US := (VAL_US)^2]
    subdat[, zerrec:= (VAL <= 0) + 0]
    subdat[, recs:= 1]
    subdat[, zblank := ifelse(MFlag == 'P', 1, 0)]
    subdat[, zerobs:= (zerrec - zblank)]
    subdat[, obs := (recs - zblank)]

    # Include additional week labels for summarizing by week (already have year and month)
    subdatwk <- as.data.table(left_join(subdat, weeklabel, by = 'date'))
    rm(subdat)

    # Select columns to summarize based on SelElem
    wkcol <- names(subdatwk)[c(4, 13:19)]   #(*Change columns)

    weeksum <- subdatwk[, lapply(.SD, sum, na.rm=TRUE), .SDcols=wkcol,
                     by=list(StationID, loc, elem, weekno, yrgrp, ctrweek)]
    setnames(weeksum, 'VAL', 'VALw')
    setnames(weeksum, 'VAL_US', 'VALw_US')
    firstcols <- weeksum[, StationID:VALsqd_US]
    firstcols[, VALsqw_US := (VALw_US)^2]    # PRCP 3,339 rows = 64 States x 52 weeks roughly
    weeksum <- cbind(firstcols, weeksum[, zerrec:obs])

    rm(firstcols)
    rm(subdatwk)
    rm(wkcol)

    # Sort columns
    weeksum <- weeksum[order(elem, loc, weekno),]

    if(q == 1){
      weeksall <- weeksum
    }
    if( q > 1){
      weeksall <- rbind(weeksall, weeksum)
    }

    rm(weeksum)

    gc()  # Call for garbage can to spare memory

  }

  ########################################
  #     END FIRST INNER LOOP (Years)     #
  ########################################

  rm(daily, csvfiles, q)

  # Weeks numbered according to Nino indices overlap years;
  # Sum again to complete summation of boundary weeks.
  wkscol <- names(weeksall)[7:15]  #(*Change columns)
  weekly <- weeksall[, lapply(.SD, sum, na.rm=TRUE), .SDcols=wkscol,
                      by=list(StationID, loc, elem, weekno, yrgrp, ctrweek)]
  rm(weeksall)  #PRCP sum removes 306,156 rows
  rm(wkscol)
  # Calculate missing or closed records based on complete 7-day week sums
  weekly[, misclsd := (7 - obs)]

  gc()


  # Read in list of stations with state/province, elevation, coordinates
  setwd(dirbase)
  stnlist <- read_csv('ghcnd-stations.csv', col_names = TRUE)
  colnames(stnlist) <- c("StationID","lat", "lon", "elev", "St", "Name", "GSNFlag", "zip" )
  stnlist$loc <- as.character(substring(stnlist$StationID, 1,2))
  stnlist$StationID <- as.character(stnlist$StationID)
  stnsub <- stnlist[,c(1:5,9)]  #(*Change columns*)
  USCANstn <- subset(stnsub, stnsub$loc %in% c('US', 'CA'))
  rm(stnlist, stnsub)

  # Read in Nino Indices
  setwd(dirbase)
  ninoweekly <- read_csv('NinoWeekly.csv', col_names = FALSE)
  colnames(ninoweekly) <- c("ctrweek","Nino12Ind", "Nino12Anom", "Nino3Ind", "Nino3Anom",
                             "Nino34Ind", "Nino34Anom", "Nino4Ind", "Nino4Anom")


  ########################################
  #   BEGIN SECOND INNER LOOP (Years)    #
  ########################################
  # Select data year by year and merge with station locations and weekly Nino indices
  for  (p in 1:length(unique(weekly$yrgrp))){
    yrwrite <-  as.integer(min(weekly$yrgrp)) + (p-1)
    weeksel <- weekly[yrgrp == yrwrite]

    weekst <- as.data.table(left_join(weeksel, USCANstn[,1:5], by = 'StationID'))
    rm(weeksel)
    weekselst <- cbind(weekst[, StationID:elem], weekst[,.(St)], weekst[,lat:elev],
                      weekst[,weekno:misclsd])   #(*Change columns*)
    rm(weekst)
```

```
    # Include weekly Nino Indices from 1990 to present in weekly table
    weekselIndex <- as.data.table(left_join(weekselst, ninoweekly, by = 'ctrweek'))
    rm(weekselst)

    setwd(diroutput)
    filenmweek <- paste0("USCANweek", Elements[z], yrwrite, "df.csv")
    write_csv(weekselIndex, filenmweek, col_names = TRUE)
    rm(yrwrite, weekselIndex)
  }

  ########################################
  #    END SECOND INNER LOOP (Years)     #
  ########################################
  rm(USCANstn, ninoweekly, weekly, filenmweek)

}
########################################
#     END OUTER LOOP (Elements)        #
########################################

rm(p, z, weeklabel)

##### End Program Code


# EXTRA CODE - does not link to weekly indices

# ------- weekly labels for weeks of all years to fall on the same days -----------
# Choose any non-leap year from Jan 1 to Dec 31
BegDate <- "1961/01/01"
EndDate <- "1961/12/31"
dt = seq(as.Date(BegDate), as.Date(EndDate), "days")  # sequence of dates from beg to end
weekno = as.integer((c(rep(1:52, each = 7), 52)))

startlab <- as.Date(BegDate)+3
centerdates <- seq(as.Date(startlab), by = "7 days", length.out = 52)
ctrweek <- c(rep(centerdates, each = 7), tail(centerdates, 1))

weeklab <- data.frame(weekno, ctrweek, dt)
dtleap = as.Date("1964/02/29")  # Choose any leap day
newrow <- data.frame(weekno = as.integer(9), ctrweek = centerdates[9], dt = dtleap)
weeklabel <- rbind(weeklab, newrow)
weeklabel$moday <- format(as.Date(weeklabel$dt), "%m-%d")
weeklabel$mdchar <- as.character(paste0(substring(as.character(weeklabel$moday), 1, 2),
substring(as.character(weeklabel$moday), 4, 5)))
weeklabel <- weeklabel[order(weeklabel$moday),]
weeklabel <- weeklabel[,c(1,5, 2)]
rm(BegDate, EndDate, centerdates, dt, dtdleap, newrow, ctrweek, startlab, weeklab)

# Code to advance year (if merging by full date)
weeklabel$ctrweek <- paste0(as.character(as.integer(substring(weeklabel$ctrweek, 1, 4))+1),
                            substring(weeklabel$ctrweek, 5, 10))

# =========================================================================================== #
# ===== CODE 8 ============= SELECT STATIONS and SUMMARIZE BY STATE ========================= #
# =========================================================================================== #
# =================================== PLOT STATIONS ========================================= #
# =========================================================================================== #
# ================================ MAP STATE CHOROPLETH ===================================== #
# =========================================================================================== #
# Go back to repeat SETUP at top if R has been closed.

# Select base years (eg. 1961 - 1990) for climatology, typically 30 past years
BegBsYr <- 1961
EndBsYr <- 1990

# Select weather element to identify summary files
SelElem <- 'PRCP'

# for mapping - read in state names and abbreviations
setwd(dirbase)
stabbr <- read_csv('ghcnd-states.csv', col_names = FALSE)
colnames(stabbr) <- c('St', 'Name')
stabbr$Name <- tolower(stabbr$Name)  # need lower case for package 'maps'

# Set working directory to access output of yearly csv files
setwd(diroutput)
yrfilenm <- paste0('_USCANyrinvgrid', SelElem, 'df.csv')
yrall <- read_csv(yrfilenm, col_names = TRUE)
yrall <- as.data.table(yrall)      #   785,941 PRCP

rm(yrfilenm)

table(yrall$bryrct)
table(yrall$bsyrct)
# Select Stations from yearly data
yrsel <- yrall[loc == 'US' & bryrct == 57]
#yrsel <- yrall[loc == 'US' & bsyrct > 24 & rcyrct > 22]

# OPTION: Plot comparisons - all US stations
library(maps)
USall <- yrall[loc == 'US']
map("state")
points(USall$lon, USall$lat, pch=19, cex = 0.05, col = 'dodgerblue3')
# Plot comparisons - only selected US stations
map("state")
points(yrsel$lon, yrsel$lat, pch=19, cex = 0.05, col = 'dodgerblue3')

# REVIEW SELECTION OF STATIONS
# view count of stations in summary data
length(unique(yrall$StationID))
# view count of stations in selection
length(unique(yrsel$StationID))
# view number of stations selected by State
table(yrsel$St)

# Add field for station count by state or year count by station
yrsel[, ct := 1]
yrselcol <- c( "VALy", "VALy_US", "obs", "ct")  #(*Change columns*)
yrselsum <- yrsel[, lapply(.SD, sum, na.rm=TRUE), .SDcols=yrselcol,
                  by=list(loc, St, elem, year)]
yrstns <-   yrsel[, lapply(.SD, sum, na.rm=TRUE), .SDcols='ct',
                  by=list(StationID, St, lat, lon)]
yrselsum <- as.data.table(yrselsum)
setnames(yrselsum, 'ct', 'stnct')
yrselsum[, VALsqSt_US := (VALy_US)^2] # 56 years x 50 states
nrow(yrselsum)  # 58 years x 50 states = 2900

# Base Years State Level Summary
baseyrs <- yrselsum[year >= BegBsYr  & year <= EndBsYr]
# Calculate mean and stdev stats manually by formula
```

```
basesum <- baseyrs[, lapply(.SD, sum, na.rm=TRUE),
                   .SDcols=c('VALy_US', 'VALsqSt_US', 'stnct', 'obs'),
                   by=list(loc, St, elem)]
basesum <- as.data.table(basesum)
basesum[, bsMean := (VALy_US/30)]
basesum[, bsSD := sqrt((VALsqSt_US - ((bsMean^2) * 30))/29)]
setnames(basesum, 'stnct', 'bsstnyrs')
setnames(basesum, 'obs', 'bsstnobs')

# Use package dplyr to calculate mean and st dev, remove NA's
# Warning: dplyr summarise function will not group if package 'plyr' is loaded
# detach(package:plyr)
  anomaly <- baseyrs %>%
  group_by(loc, St, elem) %>%
  summarise(bsstnobs=sum(obs), bsstnyrs = sum(stnct), bsmean = mean(VALy_US),
            bsstdev = sd(VALy_US), rm.na = TRUE)

# head(anomaly)

# Merge base year statistics with yearly data set
yrstat <- as.data.table(left_join(yrselsum, anomaly,
                         by = c('loc', 'St', 'elem')))
# Calculate anomalies by year and state
yrstat[, anom :=  (VALy_US - bsmean) / bsstdev]

# Merge state names with yearly data set, for mapping
yrstatst <- as.data.table(left_join(yrstat, stabbr, by = 'St'))
nrow(yrstatst)

# Create a file name to describe data output to write
fileyrst <- paste0('USyrstat', SelElem, 'df.csv')
write_csv(yrstatst, fileyrst, col_names = TRUE)

# To reopen the data table
# fileyrst <- paste0('yrstat', SelElem, 'df.csv')
# yrstatst <- read_csv('yrselstatsPRCP56.csv', col_names = TRUE)
# yrstatst <- as.data.table(yrstatst)

#.........................................
#...... Choropleth from Scratch ............
#.........................................

#Load package 'maps' to plot custom choropleth
library(maps)

# Select a year of data
SelYr <- 2016
yrst <- yrstatst[year == SelYr]

# Create breaks for ranges in the standard deviations
yrst$STDEVgrp <- cut(yrst$anom,
                     breaks = c(-Inf,  -1.5,  -1,  -0.5,  0, 0.5,  1,  1.5,  2, 2.5, Inf),
                     labels = c("neg2&-", "neg1.5", "neg1.0", "neg0.5", "pos0.50",
                                "pos1.00", "pos1.5", "pos2.0", "pos2.5", "pos3&+"),
                     right = FALSE)

# Create labels and choose color scheme to match the ranges
STDEVgrp= c("neg2&-", "neg1.5", "neg1.0", "neg0.5",
            "pos0.50", "pos1.00", "pos1.5", "pos2.0", "pos2.5", "pos3&+")
colreg <- c("wheat3", "wheat2", "wheat1",  "lightyellow1", "lightcyan1", "lightskyblue1",
            "skyblue2", "skyblue3", "skyblue4", "midnightblue")
```

```
# Combine group and colors in a data frame
colregdf <- data.frame(STDEVgrp, colreg)

# Merge colors with year data
yrstcolor <- as.data.table(left_join(yrst, colregdf, by = 'STDEVgrp'))
# Remove Alaska and Hawaii to avoid error in plotting 48 states (Check DC, Puerto Rico etc)
yrstplot <- yrstcolor[St != 'AK' & St != 'HI']

# Split data table into segments by regional colors to plot colors
yrstseg <- split(data.frame(yrstplot), yrstplot[, colreg])

# Draw states then add colors for each split in a loop
map("state")
for (r in 1:length(yrstseg)){
  map("state", region=yrstseg[[r]]$Name, interior=F, fill=T, boundary=T,
      col = as.character(yrstseg[[r]]$colreg[1]), add=T)
}

# Indicate font sizes for map
par(ps = 12, cex = 0.8, cex.main = 2.2)
title(paste0("Precipitation Anomalies by State - ", SelYr))
# Add points for the selected station locations
legendtxt <- c("< -1.5", "-1.5 to -1.0", "-1.0 to -0.5", "-0.5 to 0", "0 to +0.50",
               "+0.5 to +1.0", "+1.0 to +1.5", "+1.5 to +2.0", "+2.0 to +2.5", "> +2.5")
par(ps = 16)
legend("bottomright", legendtxt,  horiz = FALSE, fill = colreg)
# Add points for station locations
points(yrstns$lon, yrstns$lat, pch = 19, cex = 0.3, col = 'maroon4')


#.................................................
#....... Choropleth by Package ggplot2 ............
#.................................................

library(ggplot2)
library(fiftystater)

# ggplot2 base choropleth colors (data can be yrst or yrstcolor)

t <- ggplot(yrstcolor, aes(map_id = yrstcolor$Name, fill=yrstcolor$anom)) +
  geom_map(map = fifty_states, colour = 'black') +
  expand_limits(x = fifty_states$long, y = fifty_states$lat) +
  coord_map() + ggtitle(paste0('Precipitation Anomalies by State in ',
                               SelYr, '\n(Base Years ', BegBsYr,  ' - ', EndBsYr, ')')) +
  theme(plot.title=element_text(size = rel(1.5), lineheight = .9,
                                family = 'Times', colour = 'black', hjust = 0.5)) +
  theme(axis.title.x=element_blank())+
  theme(axis.title.y=element_blank())

t + fifty_states_inset_boxes()

# assigning custom choropleth colors centere at zero

q <-   ggplot(yrstcolor, aes(fill = anom, map_id = Name)) +
  geom_map(map = fifty_states, colour = 'black') +
  expand_limits(x = fifty_states$long, y = fifty_states$lat) + coord_map() +
  scale_fill_gradient2(low = "wheat4", mid = "white", midpoint = 0,
                     high = "dodgerblue4", limits = c(-3,3)) +
  ggtitle(paste0('Precipitation Anomalies by State in ', SelYr,
               '\n(Base Years ', BegBsYr,  ' - ', EndBsYr, ')')) +
  theme(plot.title=element_text(size = rel(1.5), lineheight = .9,
                                family = 'Times', colour = 'black', hjust = 0.5)) +
  theme(axis.title.x=element_blank())+
  theme(axis.title.y=element_blank())

q + fifty_states_inset_boxes()

##### End Program Code


# =========================================================================================== #
# ===== CODE 9 ================= COMBINE MONTHLY INDICES =================================== #
# =========================================================================================== #

# Set boundaries on data to combine – make sure years exist in the data
MinYr <- 1951
MaxYr <- 2017

# Read in monthly Nino Indices
setwd(dirbase)
ninomonthly <- read_csv('NinoMonthly.csv', col_names = TRUE)
nino <- as.data.table(ninomonthly)
nino
colnames(nino) <- c('year', 'month', "Nino12", "Anom12", "Nino3", "Anom3", "Nino4", "Anom4",
                      "Nino34", "Anom34")
nino
nino <- nino[year >= MinYr & year <= MaxYr]

SOI <- read_csv('SOI_Anom.csv', col_names = FALSE)
EQSOI <- read_csv('EQSOI.csv', col_names = FALSE)
TNI <- read_csv('TNI.csv', col_names = FALSE)
BEST <- read_csv('BEST1mo.csv', col_names = FALSE)
colnames(SOI) <- c('year', seq(1:12))
colnames(EQSOI) <- c('year', seq(1:12))
colnames(TNI) <- c('year', seq(1:12))
colnames(BEST) <- c('year', seq(1:12))

SOIlg <- as.data.table(gather(SOI, "month", "SOI", 2:13))
EQSOIlg <- as.data.table(gather(EQSOI, "month", "EQSOI", 2:13))
TNIlg <- as.data.table(gather(TNI, "month", "TNI", 2:13))
BESTlg <- as.data.table(gather(BEST, "month", "BEST", 2:13))
SOIlg <- SOIlg[year >= MinYr & year <= MaxYr]
EQSOIlg <- EQSOIlg[year >= MinYr & year <= MaxYr]
TNIlg <- TNIlg[year >= MinYr & year <= MaxYr]
BESTlg <-  BESTlg[year >= MinYr & year <= MaxYr]

IndexAll <- cbind(nino, SOIlg[,3], EQSOIlg[,3], BESTlg[,3], TNIlg[,3])

write_csv(IndexAll, 'IndexMonthly.csv', col_names = TRUE)


##### End Program Code


# =========================================================================================== #
# ===== CODE 10 ================= PLOT INDEX TIME SERIES ==================================== #
# =========================================================================================== #

# Load packages
library(reshape2)   # To convert Index data from wide to long
library(ggplot2)    # To produce graphs of indices
```

```
setwd(dirbase)
indices <- read_csv('IndexMonthly.csv', col_names = TRUE)
indices <- as.data.table(indices)
indices[indices < -99] <- NA
IndexName <- names(indices)[3:14] #(*Change Columns*)

IndexName  # view index names (column headers)
# Select index number
i = 4
# assign column number of selected index
indcol <- as.integer(i+2)
SelIndex <- names(indices)[indcol]
# Rename column to be plotted
setnames(indices, SelIndex, "PlotIndex")
PlotDat <- cbind(indices[, year:month], indices[, .(PlotIndex)])
PlotDat$positive <- PlotDat$PlotIndex >= 0  # TRUE/FALSE values
ggplot(PlotDat, aes(x=year, y = PlotIndex, fill = positive))  + geom_bar(stat="identity") +
ylab(SelIndex)
# Reset name of column plotted to original index name
setnames(indices, "PlotIndex", SelIndex)

# If error, reset column names
# colnames(indices) <- c(names(indices)[1:2], IndexName)

MEI <- read.csv('MEI_Index.csv', header = TRUE)
colnames(MEI) <- c('Year', 'DecJan', 'JanFeb', 'FebMar', 'MarApr', 'AprMay', 'MayJun',
                   'JunJul', 'JulAug',  'AugSep', 'SepOct', 'OctNov', 'NovDec')
head(MEI)

# Labels used to sort bimo and trimo ascending for time series
molabels <- data.frame(month = seq(1, 12, 1),
                       mo = c('Jan','Feb', 'Mar', 'Apr', 'May', 'Jun',
                              'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec'),
                       bimo = c('JanFeb','FebMar', 'MarApr', 'AprMay', 'MayJun', 'JunJul',
                                'JulAug', 'AugSep', 'SepOct', 'OctNov', 'NovDec', 'DecJan'),
                       trimo = c('JFM','FMA', 'MAM', 'AMJ', 'MJJ', 'JJA',
                                 'JAS', 'ASO', 'SON', 'OND', 'NDJ', 'DJF'))

# use tidyvers gather() to convert wide format to long
MEIlong <- gather(MEI, "bimo", "MEI", 2:13)
MEIlong[MEIlong < -99] <- NA
head(MEIlong)

# Create column to identify positie values in order to color code graph plot
MEIlong$posM <- MEIlong$MEI >= 0    # TRUE/FALSE values
MEIlongno <- merge(MEIlong, molabels[,c(1,3)], by = 'bimo', all.MEIlong = TRUE)
MEIlongno <- MEIlongno[order(MEIlongno$Year, MEIlongno$month),]
head(MEIlongno,15)

ggplot(MEIlongno, aes(x=Year, y = MEI, fill = posM)) + geom_bar(stat="identity") + ylab("MEI")

##### End Program Code


# =========================================================================================== #
# ===== CODE 11 ================ PLOT ELEMENT vs. INDEX by STATE ============================ #
# =========================================================================================== #
# Go back to repeat SETUP at top if R has been closed.
library(stringr)  # Converts all capitals to title format

# Base years (eg. 1961 - 1990) for climatology, that agree with data
```

```
BegBsYr <- 1961
EndBsYr <- 1990

Elements <- c('PRCP', 'SNOW', 'SNWD', 'TMAX', 'TMIN', 'WIND')

# SELECTIONS FOR PLOTS
SelElem <- 'SNOW'    # Four character weather element abbreviation
SelState <- 'MN'     # Two character state abbreviation
SelIndex <- 1        # 1, 2, 3, ..., 12 column order of index in IndexMonthly.csv
SelMonth <- 1        # 1, 2, 3, ..., 12 month
# Select first month of bimonthly/trimonthly sums:
# 1 = 'JanFeb'/'JFM'; 12 = 'DecJan'/'DJF'

MoLabel <- c('January', 'February', 'March', 'April', 'May', 'June', 'July',
             'August', 'September', 'October', 'November', 'December')
Month2 <- ifelse(SelMonth == 12, 1, SelMonth+1)
Month3 <- ifelse(SelMonth == 11, 1, ifelse(SelMonth == 12, 2, SelMonth +2))
SelBiMo <- paste0(substring(MoLabel[SelMonth], 1, 3), substring(MoLabel[Month2], 1, 3))
TriMoSel <- paste0(substring(MoLabel[SelMonth], 1, 1), substring(MoLabel[Month2], 1, 1),
                   substring(MoLabel[Month3], 1, 1))
BiMoLab <- paste(substring(SelBiMo, 1, 3), substring(SelBiMo, 4, 6), sep = ' - ')
TriMoLab <- paste(BiMoLab, substring(MoLabel[Month3], 1, 3), sep = ' - ')
rm(Month2, Month3)
ElemLab <- data.frame(Elem = Elements[1:5],
               ElemName = c('Rainfall', 'Snowfall','Snow Depth',
              'Max. Temperature', 'Min. Temperature'))
# for mapping - read in state names and abbreviations
setwd(dirbase)
stabbr <- read_csv('ghcnd-states.csv', col_names = FALSE)
colnames(stabbr) <- c('St', 'Name')
# read in indices
setwd(dirbase)
indexmo <- read_csv('IndexMonthly.csv', col_names = TRUE)

# Set Working Directory to access csv data files to plot
setwd(diroutput)

#  Read in Yearly Inventory
yrfile <- paste0('_USCANyrinvgrid', SelElem, 'df.csv')
yrdat <- read_csv(yrfile, col_names = TRUE)
yrdat <- as.data.table(yrdat)

# Select stations based on yearly inventory - customize code here
# Input selection criteria for stations (record completeness, etc)
selyrdat <- yrdat[bryrct == 57]   # All States
nrow(yrdat)
nrow(selyrdat)
selstn <- data.frame(StationID = unique(selyrdat$StationID), keep = 'Y')
# rm(selyrdat, yrdat)  # Remove once selections are decided
rm(yrfile)


#***********************************************#
###  CHOOSE FROM THREE DATA FILES TO READ IN  ###
#***********************************************#

# Set Working Directory to access csv data files to plot
setwd(diroutput)

# For Nino Indices, SOI, EQSOI, TNI, and 'BEST', read Year Month data
yrmofile <- paste0('_USCANyrmoinv', SelElem, 'df.csv')
```

```
yrmodat <- read_csv(yrmofile, col_names = TRUE)
yrmodat <- as.data.table(yrmodat)

# For MEI read bimonthly data
MEIfile <- paste0('_USCANbimo', SelElem, 'df.csv')
MEIdat <- read_csv(MEIfile, col_names = TRUE)
MEIdat <- as.data.table(MEIdat)

# For ONI read trimonthly data
ONIfile <- paste0('_USCANtrimo', SelElem, 'df.csv')
ONIdat <- read_csv(ONIfile, col_names = TRUE)
ONIdat <- as.data.table(ONIdat)


#**********************#
#   MONTHLY INDICES    #
#**********************#

# Keep selected stations, merged with selected data set
seldatmo <- as.data.table(left_join(yrmodat, selstn, by = 'StationID'))
nrow(seldatmo)
seldatmo <- seldatmo[keep == 'Y']
seldatmo[ , keep := NULL]
nrow(seldatmo)
seldatmo[, stns := 1]
seldatmo[,base := ifelse(year <= EndBsYr & year >=BegBsYr, 'Y', 'N' )]
seldatmo[,pre82 := ifelse(year < 1982, 'Y', 'N' )]

# Columns to sum by state for Nino, SOI, EQSOI, TNI, BEST
selmocols <- names(seldatmo)[c(14:26)]   #(*Change columns*)
Statedat <- seldatmo[, lapply(.SD, sum, na.rm=TRUE), .SDcols=selmocols,
                     by=list(St, loc, elem, year, month, pre82)]
Statedat[, avgVALm_US := (VALm_US / obs) ]
Statedatmo <- as.data.table(left_join(Statedat, indexmo, by = c('year', 'month')))
IndexName <- names(Statedatmo)[21:32]
IndexName

########## RUN PLOT ################
# Ok to change SelMonth, SelState, and SelIndex at this point.
# Optional Loop - uncomment two lines plus end bracket } to remove
for (indloop in 1:length(IndexName)){
  SelIndex <- indloop

  # Choose the index column to plot
  setnames(Statedatmo, IndexName[SelIndex], 'IndexPlot')

  # Limit data to plot selections
  Statedatplot <- Statedatmo[month == SelMonth & St == SelState & !is.na(IndexPlot)]

  # View range of values for setting y-axis limits
  min(Statedatplot$avgVALm_US)
  max(Statedatplot$avgVALm_US)
  # View range of values for setting x-axis limits
  min(Statedatplot$IndexPlot)
  max(Statedatplot$IndexPlot)

  # Define boundaries of plot
  xlo <- floor(min(Statedatplot$IndexPlot))
  xhi <- ceiling(max(Statedatplot$IndexPlot))
  yhi <- ceiling(max(Statedatplot$avgVALm_US)*100)/100
  # Select colors and point shapes
  color1 <- 'lightpink4'
  color2 <- 'darkcyan'
  pch1 = 1
  pch2 = 18


  # Labels and Title - revise as needed
  ElemLabel <- subset(ElemLab, ElemLab$Elem == SelElem)
  SelSt <- subset(stabbr, stabbr$St == SelState)
  SelStPlot <- str_to_title(SelSt[2])
  PlotTitle <- paste0(SelStPlot, " ", MoLabel[SelMonth], " ", ElemLabel$ElemName, " vs. ",
              IndexName[SelIndex],"  Index")

  # Widen Plot Region before running plot code
  plot(Statedatplot$IndexPlot, Statedatplot$avgVALm_US, xlab=paste0("Monthly ",
        IndexName[SelIndex], " Index"),
       ylab="average station measurement", xlim=c(xlo, xhi), ylim=c(0, yhi),
       main=PlotTitle,
       pch = ifelse(Statedatplot$pre82=='Y', pch1, pch2), cex.main=1.2, frame.plot=FALSE,
       col=ifelse(Statedatplot$pre82=='Y', color1, color2))
  legend(xlo, yhi, pch=c(pch1, pch2), col=c(color1, color2),
              c("prior to 1982", "1982 and on"),
              bty="o",  box.col="darkgreen", cex=.8)
  # Label outlier points with year - choose boundaries at right and left of plot
  Statedatplot[, outlier := ifelse(IndexPlot > xhi – 0.5 | IndexPlot < xlo + 0.5, year, "")]
  text(Statedatplot$IndexPlot, Statedatplot$avgVALm_US, Statedatplot$outlier, pos=1, cex=0.6)


  # Option: linear regression
  # reg<-lm(avgVALm_US~IndexPlot, data=Statedatplot)
  # abline(reg, lty =2, col = 'grey50')

  # Reset column names in data
  setnames(Statedatmo, 'IndexPlot', IndexName[SelIndex])

}

# In case of error, restore original column names
# colnames(Statedatmo) <- c(names(Statedatmo)[1:20], IndexName)

#**********************#
#         MEI          #
#**********************#

# Keep selected stations, merged with selected data set
seldat <- as.data.table(left_join(MEIdat, selstn, by = 'StationID'))
nrow(seldat)
seldat <- seldat[keep == 'Y']
seldat[ , keep := NULL]
nrow(seldat)
seldat[, stns := 1]
seldat[,base := ifelse(year <= EndBsYr & year >=BegBsYr, 'Y', 'N' )]
seldat[,pre82 := ifelse(year < 1982, 'Y', 'N' )]

# Columns to sum by state for MEI
selcols <- names(seldat)[c(13:25)]   #(*Change columns*)
StateMEI <- seldat[, lapply(.SD, sum, na.rm=TRUE), .SDcols=selcols,
                   by=list(St, loc, elem, year, ord, bimo, MEI, pre82)]
StateMEI <- as.data.table(StateMEI)
StateMEI[, avgVAL2m_US := (VAL2m_US / obs)]
```

```
rm(seldat, MEIdat)

########## RUN PLOT ################
# Ok to change SelBiMo and SelState at this point.
# Limit data to plot selections
# SelState <- 'FL'
# SelMonth <- 2
Month2 <- ifelse(SelMonth == 12, 1, SelMonth+1)
SelBiMo <- paste0(substring(MoLabel[SelMonth], 1, 3), substring(MoLabel[Month2], 1, 3))
StateMEIplot <- StateMEI[St == SelState & bimo == SelBiMo & !is.na(MEI)]
BiMoLab <- paste(substring(SelBiMo, 1, 3), substring(SelBiMo, 4, 6), sep = ' - ')

# View range of values for setting y-axis limits
min(StateMEIplot$avgVAL2m_US)
max(StateMEIplot$avgVAL2m_US)
# View range of values for setting x-axis limits
min(StateMEIplot$MEI)
max(StateMEIplot$MEI)

xlo <- floor(min(StateMEIplot$MEI))
xhi <- ceiling(max(StateMEIplot$MEI))
yhi <- ceiling(max(StateMEIplot$avgVAL2m_US)*100)/100

# Plot MEI graph - be sure to update ranges and title
color1 <- 'lightpink4'
color2 <- 'darkcyan'
pch1 = 1
pch2 = 18

# Labels and Title - revise as needed
ElemLabel <- subset(ElemLab, ElemLab$Elem == SelElem)
SelSt <- subset(stabbr, stabbr$St == SelState)
SelStPlot <- str_to_title(SelSt[2])
MEITitle <- paste0(SelStPlot, " ", BiMoLab, " ", ElemLabel$ElemName, " vs. MEI")


plot(StateMEIplot$MEI, StateMEIplot$avgVAL2m_US, xlab="MEI",
     ylab="average station measurement", xlim=c(xlo, xhi), ylim=c(0, yhi),
     main=MEITitle,
     pch = ifelse(StateMEIplot$pre82=='Y', pch1, pch2), cex.main=1.2, frame.plot=FALSE,
     col=ifelse(StateMEIplot$pre82=='Y', color1, color2))
legend(xlo, yhi, pch=c(pch1, pch2), col=c(color1, color2), c("prior to 1982", "1982 and on"),
       bty="o",  box.col="darkgreen", cex=.8)
# Label outlier points - choose boundaries
StateMEIplot[, outlier := ifelse(MEI > 1.6 | MEI < -1.6, year, "")]
text(StateMEIplot$MEI, StateMEIplot$avgVAL2m_US, StateMEIplot$outlier, pos=1, cex=0.6)

# Option: linear regression
reg<-lm(avgVAL2m_US~MEI, data=StateMEIplot)
abline(reg, lty =2, col = 'grey50')


##### End Program Code


# ========================================================================================== #
# ===== CODE 12 ===================== MAP ENSO INDEX REGIONS ============================== #
# ========================================================================================== #

library(maps)
library(mapproj)  # coordinate grids
```

```
# Pacific-centric Coordinates

# Nino 1+2
CoordPCNino12 <- data.frame(
  lat = c(0, 0, -10, -10, 0),
  lon = c(270, 280, 280, 270, 270)
)

# Nino 3
CoordPCNino3 <- data.frame(
  lat = c(5, 5, -5, -5, 5),
  lon = c(210, 270, 270, 210, 210)
)

# Nino 3.4
CoordPCNino34 <- data.frame(
  lat = c(5, 5, -5, -5, 5),
  lon = c(190, 240, 240, 190, 190)
)

# Nino 4
CoordPCNino4 <- data.frame(
  lat = c(5, 5, -5, -5, 5),
  lon = c(160, 210, 210, 160, 160)
)

# Equatorial SOI - West (5°N-5°S, 220°W-270°W)
CoordPCEQSOI_W <- data.frame(
  lat = c(5, 5, -5, -5, 5),
  lon = c(90, 140, 140, 90, 90)
)

# Equatorial SOI - East   (5°N-5°S, 80°W-130°W)
CoordPCEQSOI_E <- data.frame(
  lat = c(5, 5, -5, -5, 5),
  lon = c(230, 280, 280, 230, 230)
)

########################################################
# NINO INDEX REGIONS - ALL INDICES
###########################################################
# Map Pacific Centric Index SST Region
map("world2", xlim = c(80,300), ylim = c(-40, 40))

# EQSOI Regions
rect(140, -5, 90, 5, col = 'lightcyan1', border = FALSE)
rect(230, -5, 280, 5, col = 'lightcyan1', border = FALSE)

map("world2", xlim = c(80,300), ylim = c(-40, 40), add = TRUE)
map.axes()

map.grid(label = FALSE, lty = 1, col = "grey")
par(ps = 12)
title("El Nino Southern Oscillation Index Regions", family='Times')

par(ps = 10)
lines(x = CoordPCNino12$lon, y = CoordPCNino12$lat, col = "black", lwd = 2)
text(275, -3, "Nino",  family='Times')
text(275, -7, "1+2" ,  family='Times')
```

```
par(ps = 12)
lines(x = CoordPCNino4$lon, y = CoordPCNino4$lat, col = "black", lwd = 2)
text(175, 1, "NINO 4",  family='Times')
lines(x = CoordPCNino3$lon, y = CoordPCNino3$lat, col = "black", lwd = 2)
text(255, 1, "NINO 3", family='Times')
Nino34col <- 'dodgerblue3'
lines(x = CoordPCNino34$lon, y = CoordPCNino34$lat, col = Nino34col, lty = 3, lwd = 3)
text(213, 9, "NINO 3.4 / ONI", col = Nino34col, family='Times', lwd = 3)
#text(213, 9, "NINO 3.4 / ONI / 'BEST'", col = Nino34col, family='Times', lwd = 3)

SOIcol <- 'midnightblue'
DarwinPC <- c( 130.8456, -12.4634)
points(130.8456, -12.4634, cex = 1, col = SOIcol, pch = 19)
par(ps = 8.5)
text(124, -12, "Darwin", family='Times')
Tahiti <- c(210.574, -17.6509)
points(210.574, -17.6509, cex = 1, col = SOIcol, pch = 19)
par(ps = 8.5)
text(205, -15, "Tahiti", family='Times')

par(ps = 12)
lines(x = c(130.8456, 210.574), y = c(-27, -27), col = SOIcol, lwd = 1)
text(170, -30, "SOI", family='Times', lwd = 2)
lines(x = c(130.8456, 130.8456), y = c(-15, -27), col = SOIcol, lwd = 1)
lines(x = c(210.574, 210.574), y = c(-20, -27), col = SOIcol, lwd = 1)

EQSOIcol <- 'cyan4'
TNIcol <- 'mediumpurple4'
par(ps = 12)
text(150, 23, "EQSOI",  family='Times', lwd = 2, col = EQSOIcol)
text(231, 23, "EQSOI",  family='Times', lwd = 2,  col = EQSOIcol)
#text(170, 32, "TNI",  family='Times', lwd = 2, col = TNIcol)
#text(275, -27, "TNI",  family='Times', lwd = 2, col = TNIcol)
par(ps = 9)
text(150, 18, "(Western)",  family='Times', lwd = 1, col = EQSOIcol)
text(231, 18, "(Eastern)",  family='Times', lwd = 1, col = EQSOIcol)
#text(170, 27, "(Western)",  family='Times', lwd = 1, col = TNIcol)
#text(275, -32, "(Eastern)",  family='Times', lwd = 1, col = TNIcol)
arrows(150, 15, 135, 2, length = 0.1, angle = 20, col = EQSOIcol, lwd = 1.8)
arrows(231, 15, 245, -2, length = 0.1, angle = 20, col = EQSOIcol, lwd = 1.8)
#arrows(170, 23, 170, 5, length = 0.1, angle = 20, col = TNIcol, lwd = 1.8)
#arrows(275, -24, 275, -10, length = 0.1, angle = 20, col = TNIcol, lwd = 1.8)



##### End Program Code


# ========================================================================================== #
# ===== CODE 13 ====================== COSTLIEST STORMS ==================================== #
# ========================================================================================== #

# Go back to repeat SETUP at top if R has been closed.

# Read in data on Costliest Atlantic Hurricanes
setwd(dirbase)
Costly <- read_csv('CostlyStorms.csv', col_names = FALSE)
Costly <- as.data.table(Costly)
colnames(Costly) <- c("Name", "Cat", "Dmg_USB", "year", "YrNo", "BegDay", "EndDay")
Costly[, Label := paste(year, YrNo, sep = '_')]
# Create columns for formating plot labels
```

```
Costly[, textadj := 0]
Costly[, NameLab := ifelse(Costly$Dmg_USB >=7, Costly$Name, "")]
head(Costly, 10)

head(Costly, 10)
tail(Costly, 10)
#______________________________________________
# Plot All Storms
#______________________________________________
# Sort chronologically
Costly <- Costly[order(year, BegDay),]
Costly$NameLab <- ifelse(Costly$Dmg_USB >=10, Costly$Name, "")
Costly[Name == 'Hugo', 10] <- "Hugo" # Label Hugo
# Adjust text positions and eliminate some names
Costly[Name == 'Hugo', 9] <- - 0.5
Costly[Name == 'Maria', 9] <- 0.5
Costly[Name == 'Charley', 9] <-  -1.8
Costly[Name == 'Wilma', 9] <-  1
Costly[NameLab == 'Matthew', 10] <- ""
Costly[NameLab == 'Rita', 10] <- ""
Costly[NameLab == 'Irma', 10] <-  ""

# Create Bar Plot of all storms (Show in wide plots screen)
j <- barplot(Costly$Dmg_USB, ylim = c(0, 150), col = "darkblue", cex.main = 1.5,
       cex.axis = 0.8, cex.names = 0.7, names.arg = Costly$Label, las = 2,
       ylab = "U.S. Dollars ($ Billions)", xlab = "Year / Storm Number")
j
text(j + Costly$textadj, Costly$Dmg_USB+6, Costly$NameLab, cex = 1.2)
lines(x = c(0, 16), y = c(9.47, 9.47), col = "darkblue", lty = 2, lwd = 1)
text(59, 98, "Irma", cex = 1.2)
arrows(59.5, 93.5, 65.8, 68, col = "black", length = 0.1, angle = 20, lwd = 1.9)

#______________________________________________
# Storms Up to Andrew
#______________________________________________

# Select years and sort chronologically
CostlyA <- Costly[year<=1992,]
CostlyA <- CostlyA[order(year, BegDay),]
head(CostlyA)

# Adjust text positions and eliminate some names
CostlyA[Name == 'Andrew', 9] <-  -1

# Create bar plot of storms (Show in narrow plots screen)
a <- barplot(CostlyA$Dmg_USB, ylim = c(0, 30), col = "darkblue", cex.main = .8,
             cex.axis = .8, cex.names = 0.6, names.arg = CostlyA$Label, las = 2,
             ylab = "U.S. Dollars ($ Billions)", xlab = "Year / Storm Number")
a
text(a + CostlyA$textadj, CostlyA$Dmg_USB+0.9, CostlyA$NameLab, cex = 1)
lines(x = c(0, 16), y = c(9.47, 9.47), col = "darkblue", lty = 2, lwd = 1)


##### End Program Code
```

Worldwide Tropical Cyclone Activity Measured Using the Actuaries Climate Index® Methodology

Douglas J. Collins, FCAS, MAAA

Abstract: Accumulated cyclone energy (ACE) has been used by the National Oceanic and Atmospheric Administration, and others, as a measure of the strength and duration of tropical cyclones and their seasonal activity. The Actuaries Climate Index (ACI) was launched in 2016 for the U.S. and Canada, as a measure of changes in climate and coastal sea level in those countries. A component measuring tropical cyclone activity is not included in the ACI, since the low frequency of landfalling events is not suitable for a regional index. On a global basis, however, an index measuring tropical cyclone activity over land and water could be particularly useful to actuaries and those in the insurance industry with an interest in monitoring changes in tropical cyclone activity as it relates to climate change. Using the ACI methodology, this paper shows the results of a global index of tropical cyclone activity based on ACE data. Based on the index, trends in worldwide tropical cyclone activity over the period with good data (1985 – 2017) have been flat to downward, but this time period is not sufficiently long for a credible conclusion. This result is consistent with previous research by others, but no such analysis has been previously published based on the ACI methodology, in index form.

Keywords. Accumulated cyclone energy, Actuaries Climate Index®

INTRODUCTION

Accumulated Cyclone Energy is a measure of the strength and duration of tropical cyclones. ACE is calculated from the maximum estimated wind speed of a storm at 6-hour intervals over its lifetime. Specifically, accumulated cyclone energy can be calculated for any storm using the following formula,

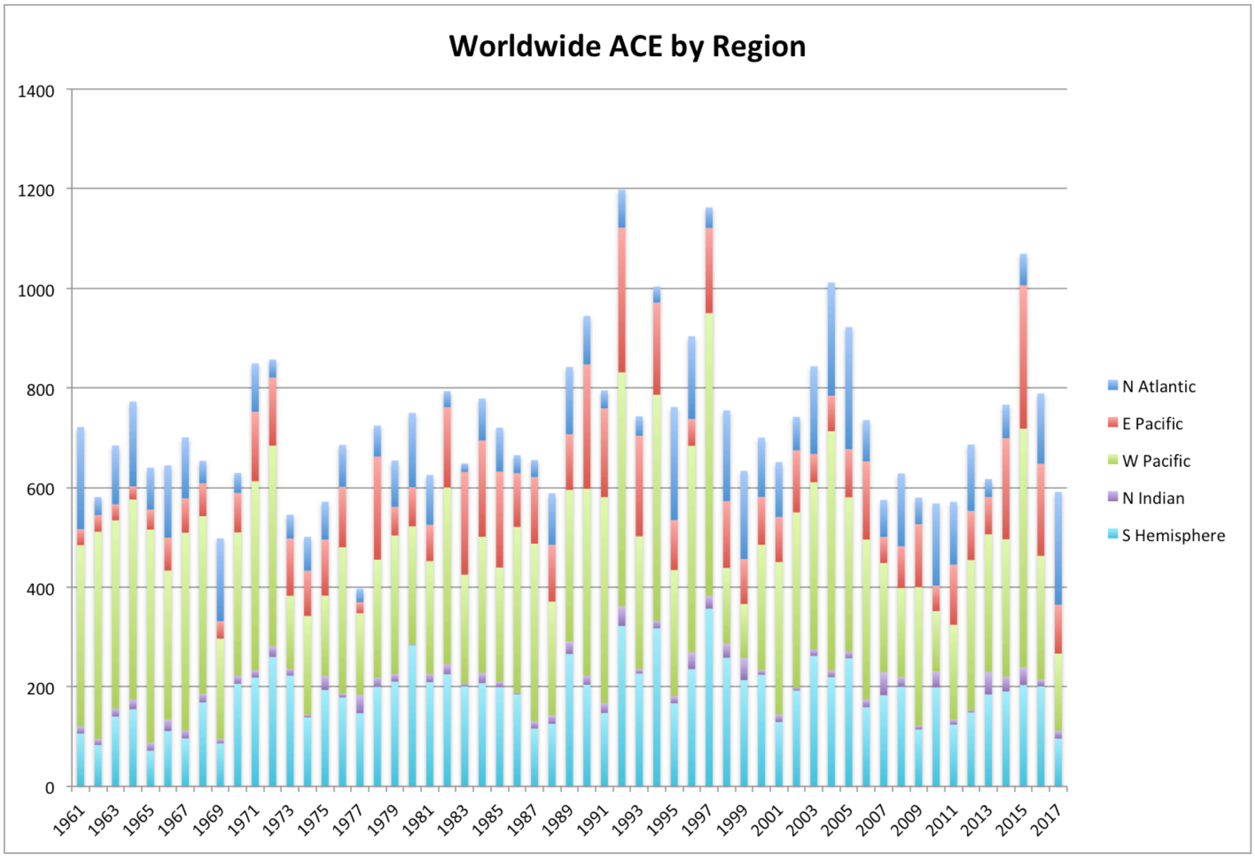
$$\text{ACE} = 10^{-4}\,\Sigma\, v^2_{\text{max}}$$

where v_{max} is the estimated maximum sustained wind speed in knots at six-hour intervals and the sum of wind speeds squared is divided by 10,000 for convenience. ACE is a more comprehensive measure than the commonly used Saffir-Simpson scale. After determining the ACE for each storm, values can be accumulated by season to provide a measure of seasonal tropical storm activity and intensity for any region of the world. A worldwide database of ACE values has been constructed by Phil Klotzbach[1],[2] of the Colorado State University Tropical Meteorology Project.

ACE data is available worldwide for most ocean basins starting in 1961 and for the Atlantic back to 1851, though tropical cyclone records are generally less reliable prior to tracking satellites, which began in the 1970s and were dependent on visible light images until the early 1980s[1]. The Actuaries Climate Index (ACI), which is documented and updated on a web site[3] sponsored by four actuarial organizations in North America (American Academy of Actuaries, Canadian Institute of Actuaries, Casualty Actuarial Society and Society of Actuaries), is calculated using a reference period of 1961-1990. As the quality of satellite data prior to the mid-1980s likely led to significant underestimates in tropical cyclone intensity[4],[5],[6], a reference period of 1985-2014 is used in this paper. While a much longer period would probably be required to measure average tropical cyclone activity, this 30-year period marks the beginning of good satellite-based data and is long enough to serve as a useful baseline.

Figure 1 shows annual ACE by region since 1961. One can see that the West Pacific is the most active basin followed by the Southern Hemisphere. In 2017, the North Atlantic basin was the largest contributor to worldwide tropical cyclone activity for the first time since at least 1961.

Figure 1 – Worldwide ACE by Region



The Actuaries Climate Index measures changes in extreme temperatures, rainfall and wind, as well as changes in sea level. The wind component in the ACI is based on average daily wind speeds and therefore does not reflect the most damaging winds found in tropical cyclones or other strong windstorms. The ACI is calculated for three-month meteorological seasons (and by month) for 12 large land regions in the United States and Canada. At this temporal and geographical scale, tropical cyclones are relatively brief occurrences that occur rarely in most of these regions and not at all in some. In order to produce a meaningful index of tropical cyclone activity, a broad geographic measure by ocean basin (including surrounding land areas) and worldwide is required. This paper summarizes such an index with annual data through 2017.

Figure 2 shows worldwide ACE as a standardized anomaly (ACE_{std}) compared to the reference period 1985-2014. The standardized anomaly, which is the metric used in the Actuaries Climate Index, is the difference between the mean ACE for each year and the reference period mean, divided by the standard deviation of ACE during the reference period. The standardized anomaly is a common tool for comparing different statistics, and measures ACE on a basis comparable to the Actuaries Climate Index. Also shown in Figure 2 are moving averages of ACE_{std} over five, ten, and twenty years. The Actuaries Climate Index is commonly presented with a five-year moving average as a means of smoothing out the random variations in the index so that trends can be more easily seen. For tropical cyclone statistics, even on a worldwide basis, five years is not long enough to accomplish similar smoothing. The ten-year average better balances stability and responsiveness and is selected as the key metric for this paper. As noted, data prior to the mid 1980s is likely understated. Over the subsequent time period, ten-year averages have been generally declining since 1998 and twenty-year averages since 2006. Research by Klotzbach[1],[5] and Maue[7],[8] has previously discussed these trends. The 21st century projections in the Fifth Assessment Report of the IPCC[9] are that the global frequency of tropical cyclones will likely decrease or remain essentially unchanged, while the global mean maximum wind speed in cyclones will likely increase.

Figure 2 – Worldwide ACE$_{std}$ using 1985-2014 Reference Period

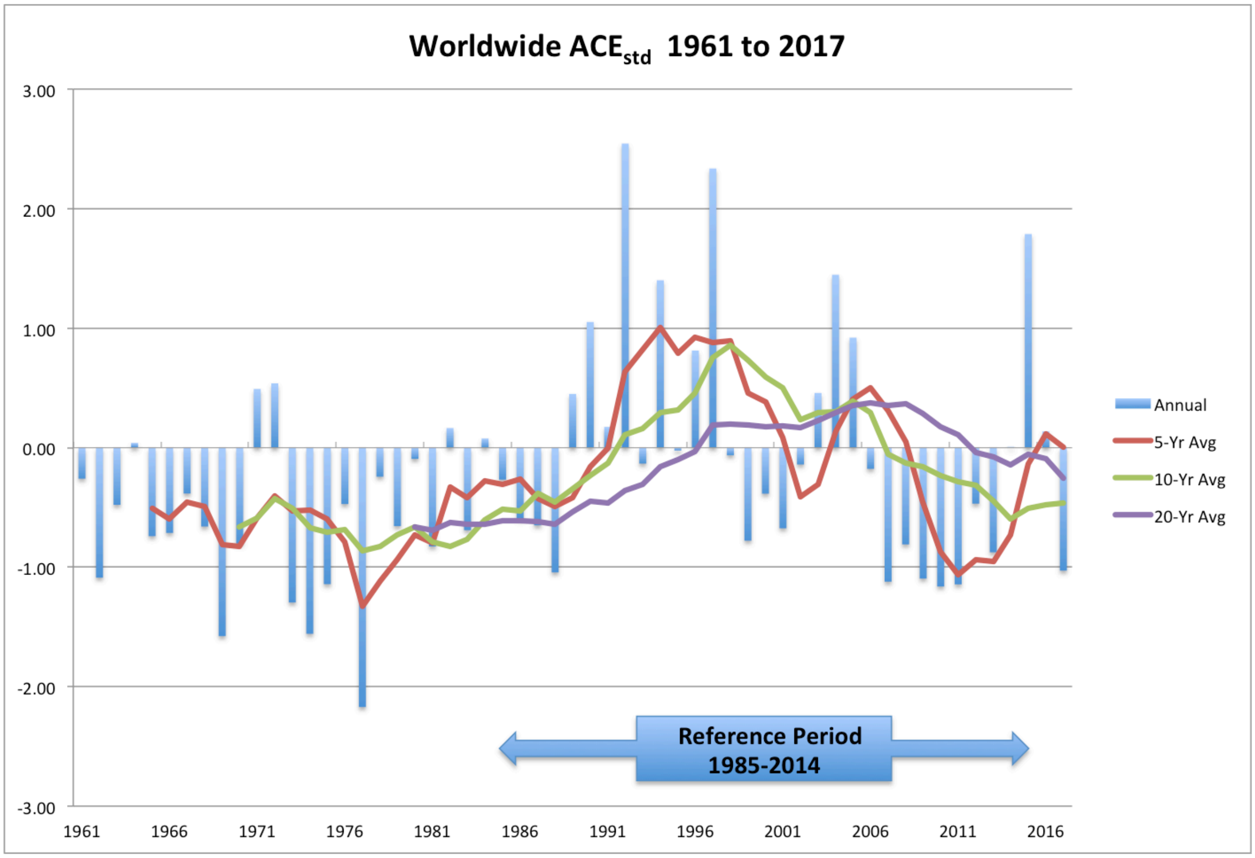
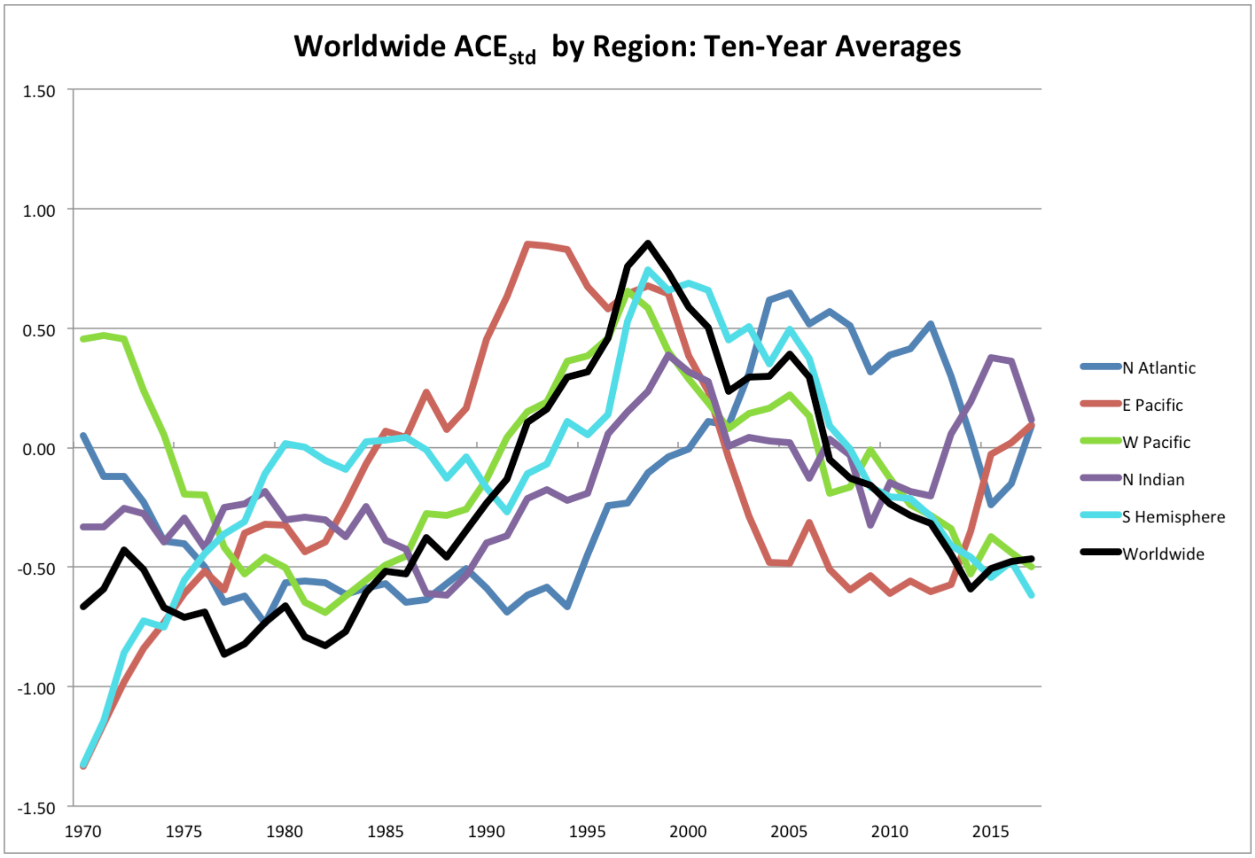


Figure 3 focuses on the ten-year moving average and compares ACE_{std} by region through 2017. One can see that activity in different regions is often offsetting. Only for the ten-year periods ending 1996-2005 have four or more of the five regions been above the reference period average. In recent years, the worldwide average has been pulled down by well-below average tropical cyclone activity in the Western Pacific and the Southern Hemisphere. Ryan Maue has noted[7] that activity in the North Atlantic and East Pacific basins has been inversely correlated and this can be clearly seen in Figure 3 (red versus dark blue lines).

Figure 3 – Comparison of regional ACE_{std} using ten-year averages



1.1 Research Context

This paper primarily focuses on the following areas from the Research Taxonomy:

- Actuarial Applications and Methodologies - Data Management and Information, Enterprise Risk Management Processes (Analyzing/Quantifying Risks), Ratemaking (Large Loss and Extreme Event Loading)

- Financial and Statistical Methods – Natural Peril Modeling
- Business Areas – Fire & Allied Lines, Homeowners, Reinsurance

A search of the Database of Actuarial Research Enquiry does not show any mention of the use of ACE to analyze hurricane activity, or any actuarial paper on global tropical cyclone or hurricane activity.

Other scientific papers on the use of ACE or analyzing global hurricane activity are:

1. Increasing destructiveness of tropical cyclones over the past 30 years, Emanuel, 2005[10]
2. Hurricanes and Global Warming, Pielke Jr., Landsea, Mayfield, Laver and Pasch, 2005[11]
3. Trends in global tropical cyclone activity over the past twenty years (1986 – 2005), Klotzbach, 2006[1]
4. A globally consistent reanalysis of hurricane variability and trends – Kossin, Knapp, Vimont, Murnane and Harper, 2007[12]
5. Northern Hemisphere tropical cyclone activity – Maue, 2009[7]
6. Recent historically low global tropical cyclone activity – Maue, 2011[8]
7. Historical Global Tropical Cyclone Landfalls – Weinkle, Maue & Pielke Jr., 2012[13]
8. Extremely intense hurricanes: revisiting Webster et al. (2005) after 10 years – Klotzbach & Landsea, 2015[5]

1.2 Objective

Given the lack of actuarial literature on the subject, and the launch in November 2016 of the Actuaries Climate Index, this paper will provide background information and data on Accumulated Cyclone Energy, in the form of an index using the ACI methodology.

1.3 Outline

In the remainder of the paper, Section 2 will discuss background and methods used, Section 3 summarizes and discusses results, and Section 4 describes results and conclusions in more detail.

2. BACKGROUND AND METHODS

2.1 ACE data

Sums of ACE statistics for all storms in a season or year provide a measure of the combined cyclone energy in that season or year. This paper is based on summaries of ACE, worldwide and by ocean basin, constructed by Phil Klotzbach and available on the Colorado State University Meteorological Project website[2]. Similar data, available on a website[14] constructed by Ryan Maue at Weather.us, was also reviewed in researching this paper and was used to fill a few gaps in the Klotzbach database.

ACE provides a more comprehensive measure of cyclone activity that the commonly used Saffir-Simpson measure of hurricane strength, which is usually cited based on the maximum wind speed over the life of a storm. Another measure, track integrated kinetic energy (TIKE), introduced by Misra et al[15], accounts for the size of the wind field, in addition to intensity and duration, making it a slightly better statistic than ACE but TIKE data is not generally available historically on a worldwide basis.

2.2 ACE statistics and values for notable storms

Exhibit 1 displays ACE statistics from 1961 through 2017 by region and worldwide. Regions are defined as follows: North Atlantic, East Pacific (north of the equator and east of 180 degrees longitude), West Pacific (north of the equator and west of 180 degrees longitude), North Indian and Southern Hemisphere. Southern Hemisphere data includes the South Indian and South Pacific and is for each year ending June 30th. Tropical cyclones are rare in the South Atlantic and are excluded from the ACE statistics. Data for each region includes the entire time that each storm is classified as either tropical or sub-tropical, whether over the ocean or surrounding land areas, bays and seas. ACE data for storms that cross regions are assigned to the region in which they first became named[2].

As noted in the footnotes on Exhibit 1, Sheet 2, the Klotzbach data goes back to 1961 or earlier (to 1851 in the North Atlantic) except in the East Pacific (EPAC) and the North Indian (NIO) regions. To fill in the early years in EPAC and NIO, which represent only 19% of worldwide ACE across all years, two sources were used:

- Maue[11] provides ACE for 1970 EPAC and 1970-1971 NIO
- Wikipedia provides ACE for EPAC in 1963, 1965 and 1966 on pages:

https://en.wikipedia.org/wiki/196n_Pacific_hurricane_season, where n is 3, 5 or 6.

- Otherwise, the author estimated 1960s ACE in these two regions based on the number and maximum intensity of storms

The largest worldwide annual value of ACE was 1,198, which occurred in 1992. On average, ACE is distributed by region as follows:

- West Pacific 41%
- Southern Hemisphere 26% (about two-thirds of which is from the South Indian, based on the Klotzbach data)
- East Pacific 16%
- North Atlantic 14%
- North Indian 3%

The largest ACE for an individual storm during the satellite era is 82 for Hurricane Ioke in 2006, which started in the East Pacific before crossing into the West Pacific and spent 9 days as a category 4 or 5 storm on the Saffir-Simpson scale. The largest ACE in the North Atlantic is 70.4 for Hurricane Ivan in 2004, which spent 8.25 days as a category 4 or 5 storm. These statistics are from Wikipedia. More recently, Hurricane Irma in 2017 generated the second highest ACE in the North Atlantic during the satellite era with a value of 67.5, spending 5.5 days as a category 4 or 5 storm according to preliminary information from the National Oceanic & Atmospheric Administration.

2.3 Applying the ACI method to ACE

The standardized anomalies of annual ACE compared to the 1985-2014 reference period are calculated based on the following formula:

$$ACE_{std} = (\text{ACE for Year N} - \text{Reference period mean ACE}) \,/\, (\text{Reference Period ACE standard deviation})$$

Exhibit 2, Sheets 1-6, show ACE_{std} worldwide and by region graphically. The underlying data for ACE_{std} worldwide is shown on Exhibit 3, Sheets 1-2 and ACE_{std} by region is shown on Exhibit 4, Sheets 1-2.

A key consideration in determining ACE_{std} is the choice of reference period. As noted, the Actuaries Climate Index uses a reference period of 1961-1990 but the more recent period was chosen for $\mathrm{ACE}_{\mathrm{std}}$ due to concerns about the completeness of tropical cyclone data prior to the advent of good satellite data.

3. RESULTS AND DISCUSSION

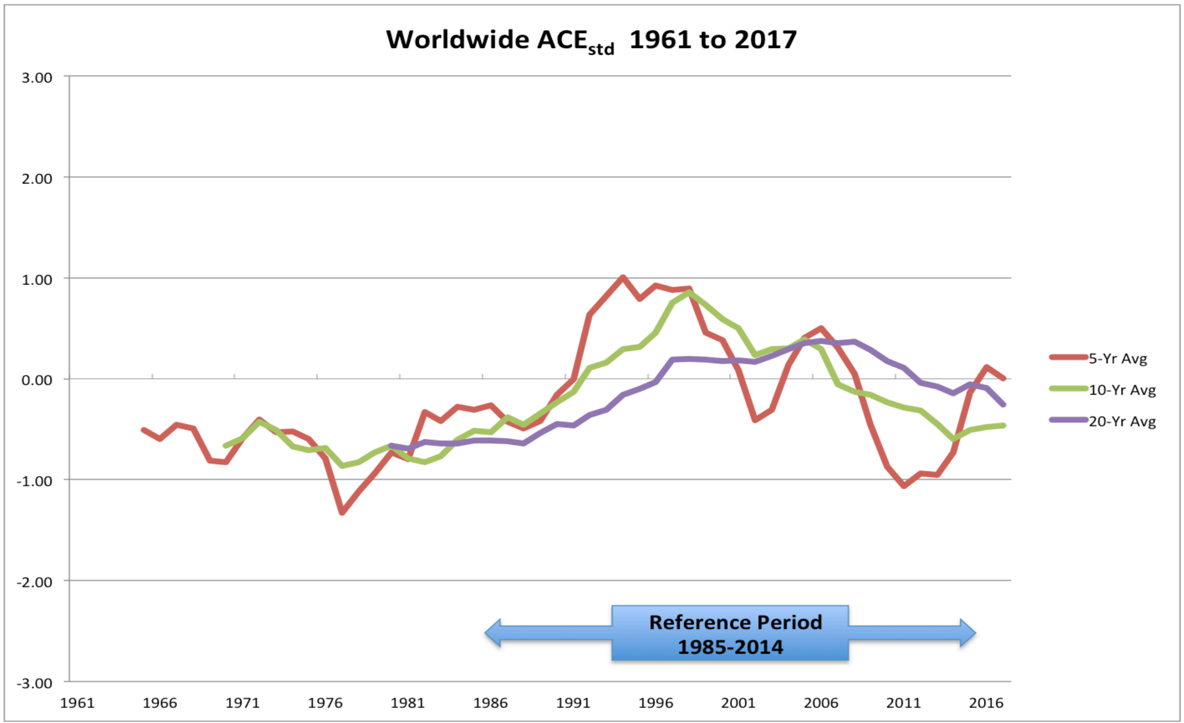
Exhibit 2, Sheets 1-6 display worldwide and regional $\mathrm{ACE}_{\mathrm{std}}$, each on a y-axis scale measuring anomalies relative to the reference period standard deviations, e.g., a value of 1.00 is a year one standard deviation above the reference period mean. Standardized anomalies are a useful way of comparing different quantities on a similar scale.

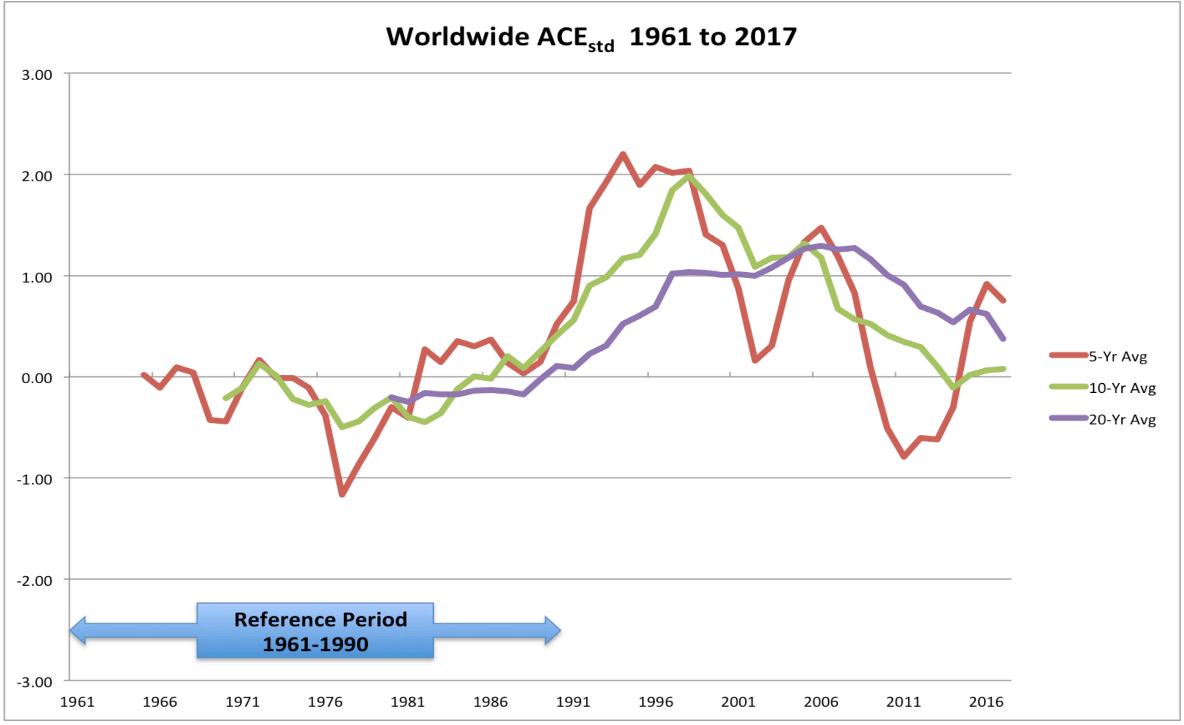
On a worldwide basis (Exhibit 2, Sheet 1), seven of the last ten years have had significantly below normal tropical cyclone activity, and only one of the last ten has had significantly above normal activity. In the North Atlantic, 2017 was one of the most active and intense years for hurricanes, but the other four ocean basins all had below normal activity bringing the worldwide total to a well-below average level.

Figure 2 shows worldwide $\mathrm{ACE}_{\mathrm{std}}$ with the 1985-2014 reference period and various rolling averages. Figure 4 below compares the rolling averages in Figure 2 with rolling averages calculated with the same reference period as the Actuaries Climate Index, i.e. 1961-1990. The underlying annual ACE data is the same in each graph but the standardized anomalies are different and can be found in Exhibit 3, Sheets 1-2. The reference period mean is lower for 1961-1990 than for 1985-2014 and the standard deviation is much lower in the earlier period. As a result, the standardized anomalies are much higher, peaking at around two in the graph based on the earlier reference period.

Whereas in Figure 2 (and the top graph in Figure 4) all three rolling averages ending 2017 were at or below the 1985-2014 reference period mean, the ten-year average ending 2017 is virtually the same as the 1961-1990 reference period mean in the lower graph in Figure 4 and the five- and twenty-year averages finish above the reference period mean. Taking the ten-year average as the best metric for measuring trends in global cyclone activity implies that even with the 1961-1990 reference period the recent rolling averages indicate no significant trend. An averaging period longer than ten years may be more appropriate for determining trends. Until much more data becomes available with good satellite measurements of tropical cyclone intensity in the coming decades, the appropriate averaging period and the existence of trends will remain uncertain given the natural variability in these statistics.

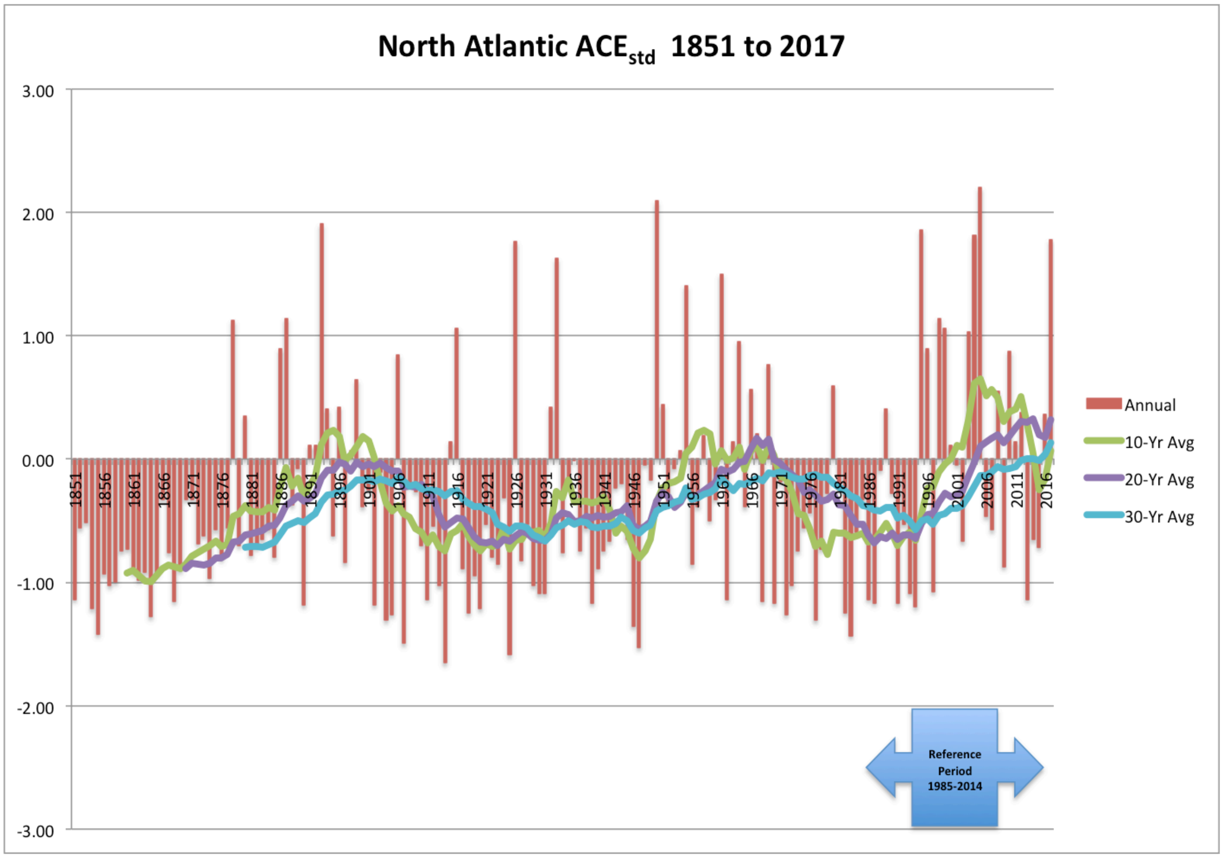
Figure 4 – Worldwide ACE_{std} Comparing Reference Periods





A longer historical time series is provided by NOAA[16] for the North Atlantic, which is probably the most-studied region but still only has high quality data back to 1966 per Klotzbach[2]. Standardized anomalies of this ACE data are plotted below in Figure 5. Note that the thirty-year averages never reached zero, i.e., the average during 1985-2014 reference period, in the pre-satellite era.

Figure 5 – North Atlantic ACE with 1985-2014 Reference Period



It has been noted by some researchers[1],[7],[8],[9],[10],[11] that although the number of global tropical cyclones has not increased with climate change, the number of intense storms has been trending upward. Data from Klotzbach[2] on storm counts is displayed in Exhibit 6, Sheets 1-5 and summarized in graphs on Exhibit 5, Sheets 1-2. This data shows that the number of named storms worldwide have been fairly stable since the 1970s in the 83-90 range per year. Ten-year averages of major storms (category 3-5) have been stable since about the mid-1990s at 23.5-25.5 per year but at a higher level than the 1970s and 1980s, consistent with research[5],[6] indicating that pre-satellite data likely led to significant underestimates in intensity, especially for the most intense systems. Exhibit 5, Sheet 2 shows a decline in the number of non-major tropical cyclones (tropical storms and category 1-2 cyclones/hurricanes/typhoons) from around 70 per year in the 1970s to 10-year averages around 60 since 2010.

4. CONCLUSIONS

The methodology presented in this paper produces an index of tropical cyclone activity worldwide and by region, consistent with the methodology used for the Actuaries Climate Index. Such an index can be easily updated and analyzed periodically.

Worldwide ACE_{std} since 1985, the period with good satellite data, is probably not long enough to credibly measure the effects of climate change on tropical cyclone activity. The two most unusual years for worldwide ACE_{std} were 1992 and 1997, the only years more than two standard deviations above the reference period mean. In the 20 years since then, there have been 14 below average years and only 6 above average years. The warm temperature component of the Actuaries Climate Index[3], and global temperature studies, have shown rapidly increasing anomalies since the late 1970s. Evidence that these warmer temperatures, along with warmer oceans, have increased the frequency and intensity of tropical cyclones remains to be seen.

Acknowledgment

The author acknowledges and appreciates the public availability of data from Ryan Maue and Phil Klotzbach, and the valuable contributions of the reviewers, Jess Geller and Alp Can.

5. REFERENCES

[1] Klotzbach, Philip J., "Trends in global tropical cyclone activity over the past twenty years (1986 – 2005)," *Geophysical Research Letters*, 2006, Vol. 33, L10805 pages 1-4.

[2] http://tropical.atmos.colostate.edu/Realtime/

[3] http://actuariesclimateindex.org/home/

[4] http://tropical.colostate.edu/real-time-cyclone-activity/#1468450487074-f84cb69f-264f

[5] Klotzbach, Philip J. and Landsea, Christopher W., "Extremely intense hurricanes: revisiting Webster et al. (2005) after 10 years", American Meteorological Society *Journal of Climate*, 2015, pages 7621-7629

[6] Hoarau, Karl et al, "Extreme tropical cyclone activities in the southern Pacific Ocean", Royal Meteorological Society *International Journal of Climatology*, 2017, 12 pages

[7] Maue, Ryan N., "Northern Hemisphere tropical cyclone activity," *Geophysical Research Letters*, 2009, Vol. 36, L05805 pages 1-5.

[8] Maue, Ryan N., "Recent historically low global tropical cyclone activity," *Geophysical Research Letters*, 2011, Vol. 38, L14803 pages 1-6.

[9] Christiansen, J.H., et al, "Climate Change 2013: The Physical Science Bases," *Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change,* 2013, Chapter 14, Section 14.6, pages 1248-1253.

[10] Emanuel, Kerry, "Increasing destructiveness of tropical cyclones over the past 30 years," letter to *nature*, 2005, Vol 436, pages 686-688.

[11] Pielke Jr., R., Landsea, C., Mayfield, M., Laver, J., and Pasch, R., "Hurricanes and Global Warming," American Meteorology Society *BAMS,* November 2005, pages 1571-1575.

[12] Kossin, J. P., Knapp, K. R., Vimont, D. J., Murnane, R. J., and Harper, B. A., "A globally consistent reanalysis of hurricane variability and trends," *Geophysical Research Letters*, 2006, Vol. 34, L104815 pages 1-6.

[13] Weinkle, J., Maue, R., and Pielke Jr., R., "Historical Global Tropical Cyclone Landfalls," American Meteorological Society *Journal of Climate,* Volume 25, 2012, pages 4729-4735.

[14] http://wx.graphics/tropical/

[15] Misra, V., DiNapoli S., and Powell M., "The Track Kinetic Energy of Atlantic Tropical Cyclones", American Meteorological Society *Monthly Weather Review,* 2013, 141:7, pages 2383-2389.

[16] https://www.esrl.noaa.gov/psd/data/timeseries/monthly/ACE/

Abbreviations and notations

ACE, Accumulated Cyclone Energy
ACI, Actuaries Climate Index
DFA, dynamic financial analysis
GLM, generalized linear models
OLS, ordinary least squares
ERM, enterprise risk management

Biography of the Author

Douglas J. Collins retired from Towers Watson in 2010 following a 30-year consulting career at TW and its predecessor firms. He began his career and completed his actuarial exams at the Travelers Insurance Company. He has a B.S. in Mathematics from Rensselaer Polytechnic Institute. He is a Fellow of the CAS and a Member of the American Academy of Actuaries. He has been a member of the Climate Change Committee since its inception and was its chair from 2014 to 2017.

Exhibit 1
Sheet 1

Accumulated Cyclone Energy (ACE)

Source: http://tropical.atmos.colostate.edu/Realtime/ (except as noted)

	North Atlantic (NATL)	East Pacific (EPAC*)	West Pacific (WPAC)	North Indian (NIO**)	Southern Hemisphere (SH***)	Worldwide (WW)
1961	205	32	364	15	106	722
1962	36	34	416	13	83	581
1963	118	32	378	16	140	684
1964	170	27	401	20	155	772
1965	84	40	428	16	72	640
1966	145	66	299	23	111	644
1967	122	69	397	16	96	700
1968	45	66	357	17	169	654
1969	166	35	202	9	86	498
1970	40	79	288	16	206	629
1971	97	139	380	15	219	849
1972	36	137	402	22	260	857
1973	48	115	147	13	222	546
1974	68	91	198	5	139	501
1975	76	113	161	28	194	572
1976	84	122	294	8	179	686
1977	27	22	164	37	147	397
1978	62	207	237	19	200	724
1979	93	57	278	15	210	654
1980	149	79	237	2	284	750
1981	100	73	227	16	209	625
1982	32	161	355	21	225	794
1983	17	206	220	4	201	648
1984	84	193	273	21	207	779
1985	88	193	229	11	199	720
1986	36	108	333	3	185	665
1987	34	134	357	15	116	655
1988	103	114	228	17	126	588
1989	135	112	304	25	266	842
1990	97	250	375	18	204	944
1991	36	178	413	20	147	795
1992	76	291	470	40	322	1,198
1993	39	202	267	9	227	743
1994	32	185	454	15	318	1,004
1995	227	100	253	14	167	762
1996	166	54	415	34	235	904
1997	41	171	568	26	357	1,162
1998	182	134	152	28	258	755
1999	177	90	109	44	213	633
2000	119	96	252	9	224	700
2001	110	91	306	15	129	651
2002	67	125	351	6	193	742
2003	176	57	335	14	262	843
2004	227	71	481	13	219	1,011
2005	245	97	310	13	257	922
2006	83	157	321	15	159	736

Exhibit 1
Sheet 2

Accumulated Cyclone Energy (ACE)

Source: http://tropical.atmos.colostate.edu/Realtime/ (except as noted)

	North	East	West	North	Southern	
	Atlantic	Pacific	Pacific	Indian	Hemisphere	Worldwide
	(NATL)	(EPAC*)	(WPAC)	(NIO**)	(SH***)	(WW)
2007	74	53	220	46	183	575
2008	146	84	178	20	200	628
2009	53	126	278	8	114	580
2010	165	51	121	32	199	568
2011	126	121	190	11	124	571
2012	133	99	302	4	148	686
2013	36	75	276	46	184	617
2014	67	203	276	30	191	766
2015	63	288	479	36	204	1,069
2016	141	185	248	14	201	789
2017	226	98	155	16	96	591
Total	5,830	6,585	17,105	1,054	10,747	41,320
Average	102	116	300	18	189	725
% of WW	14%	16%	41%	3%	26%	

Notes: ** For EPAC, 1961-2, 1964, 1967-1969 estimated by author based on Wikipedia season summaries: 196x Pacific Hurricane Season; 1963, 1965-66 from similar Wikipedia pages, which show ACE. 1970 from Maue: (http://wx.graphics/tropical/)

*** For NIO, 1961-1969 estimated by author, 1970-71 from Maue (http://wx.graphics/tropical/)

**** SH Data for 12 months ending June 30th. Excludes South Atlantic.

Exhibit 2
Sheet 1

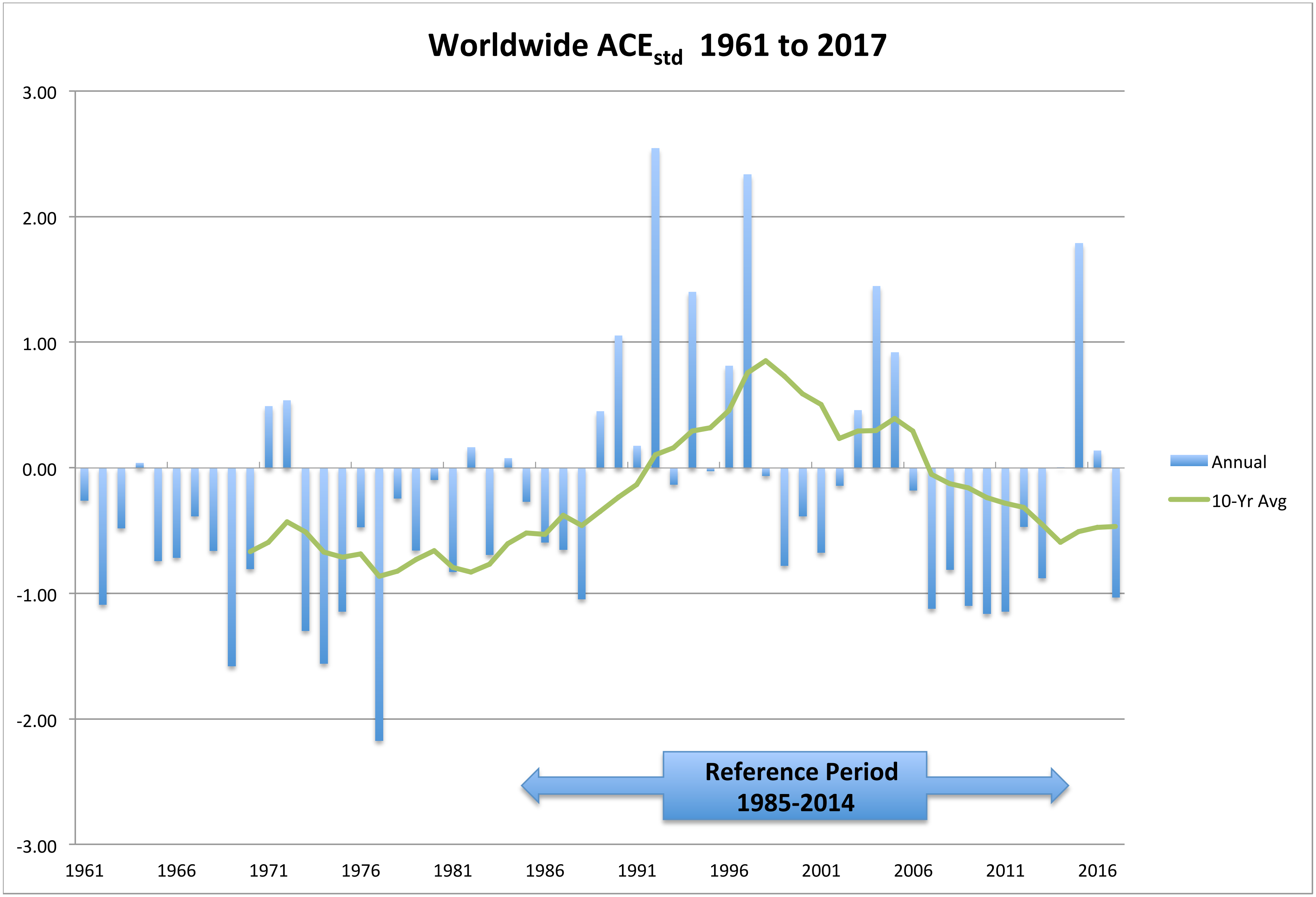


Exhibit 2
Sheet 2

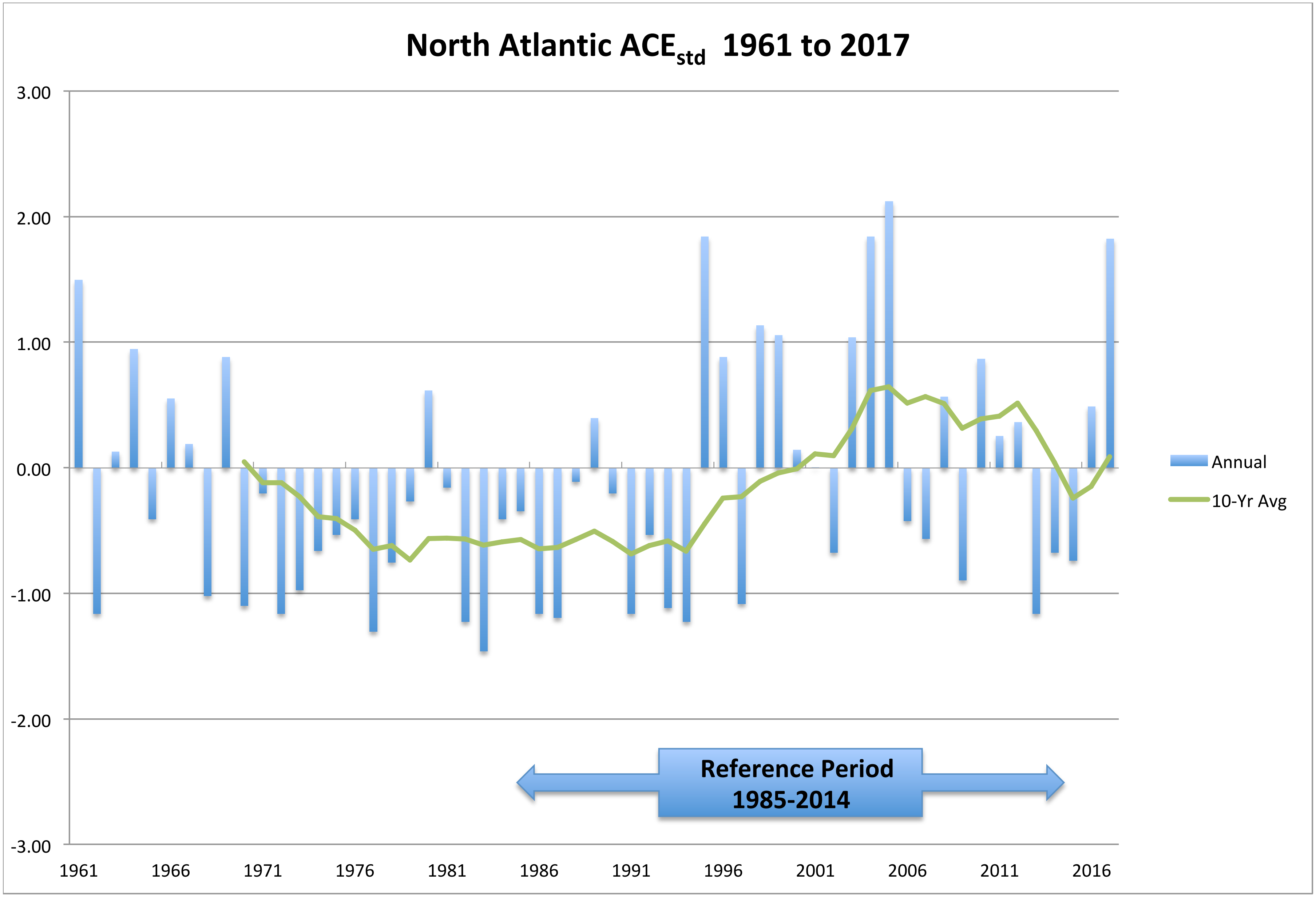


Exhibit 2
Sheet 3

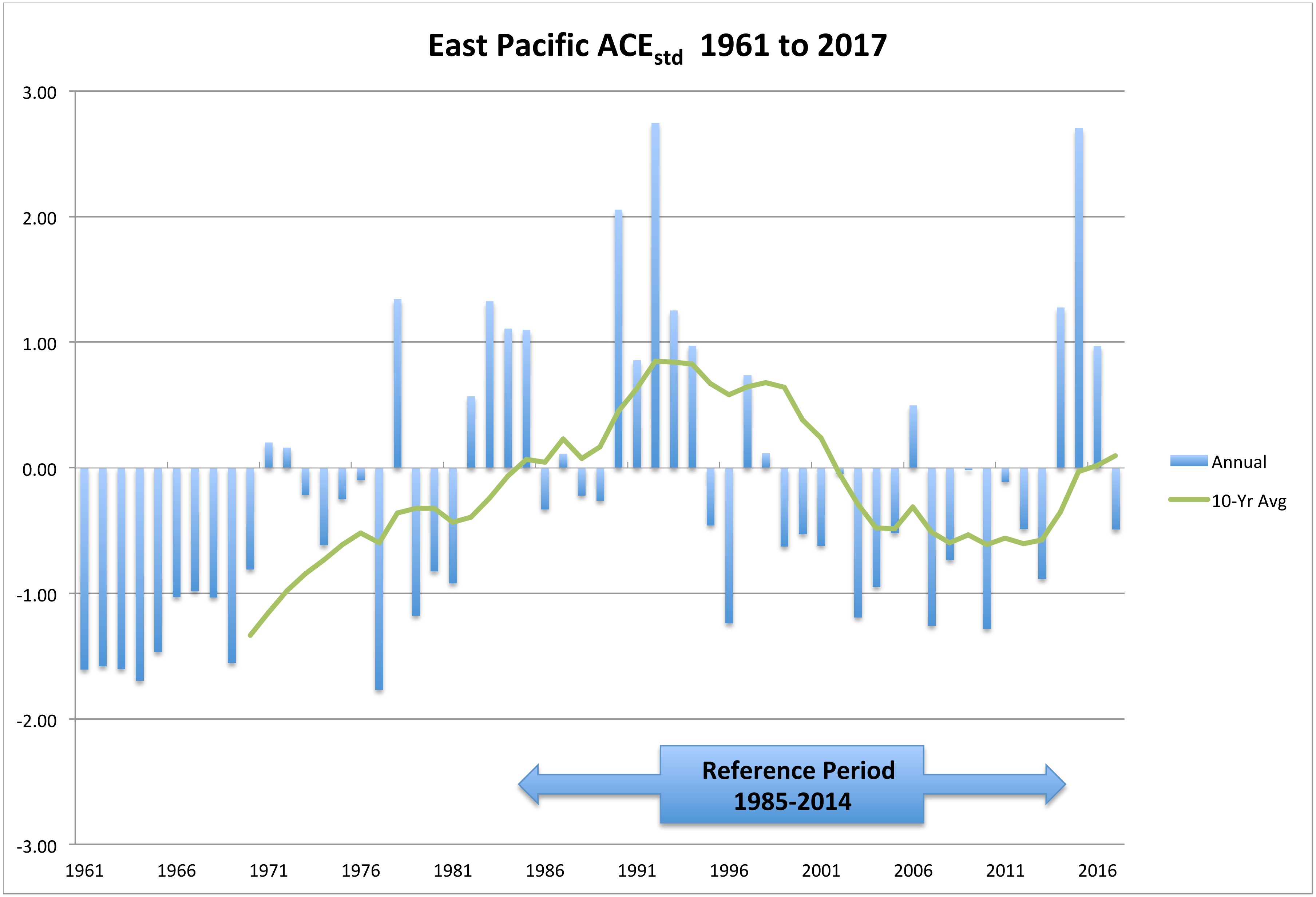


Exhibit 2
Sheet 4

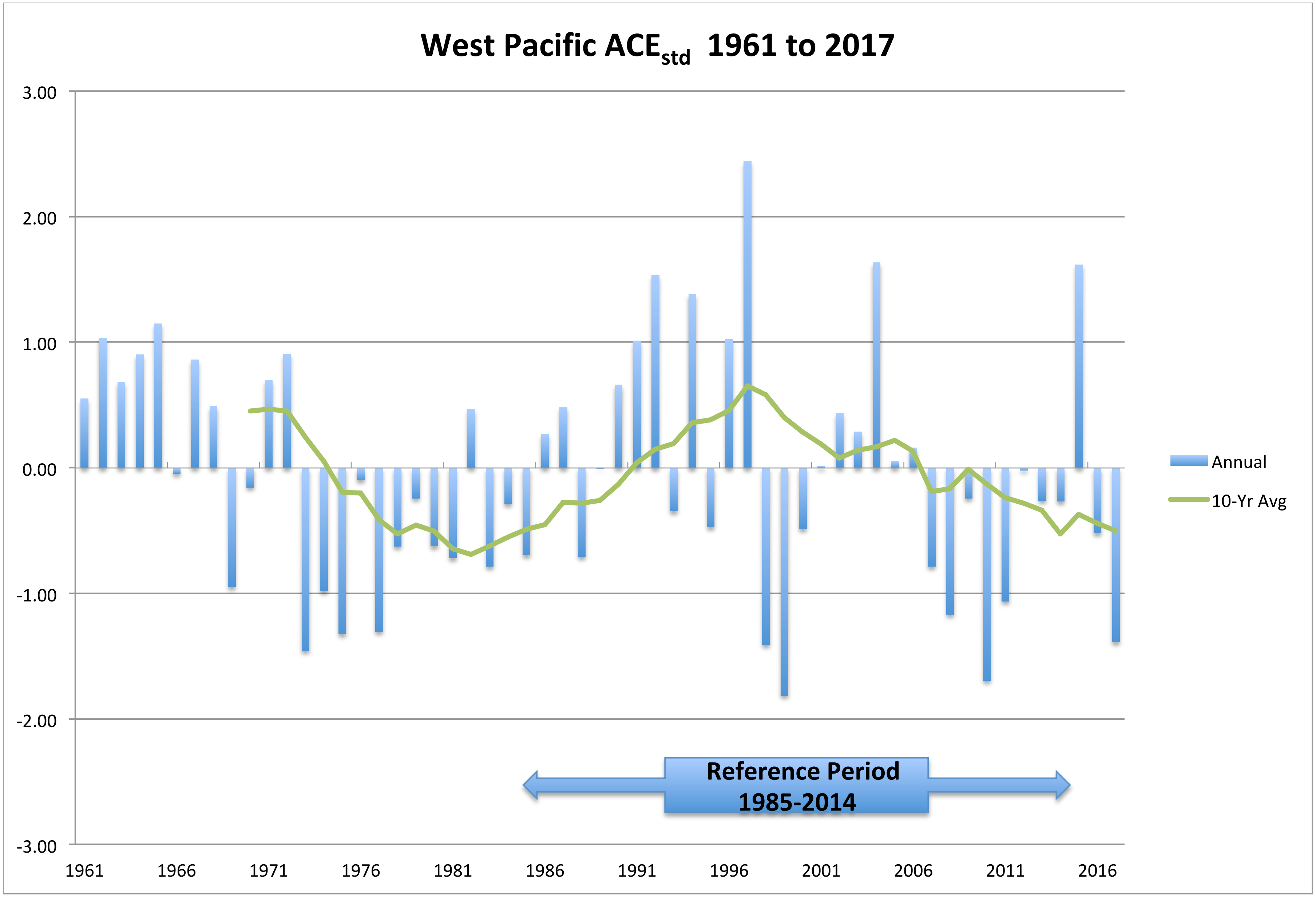


Exhibit 2
Sheet 5

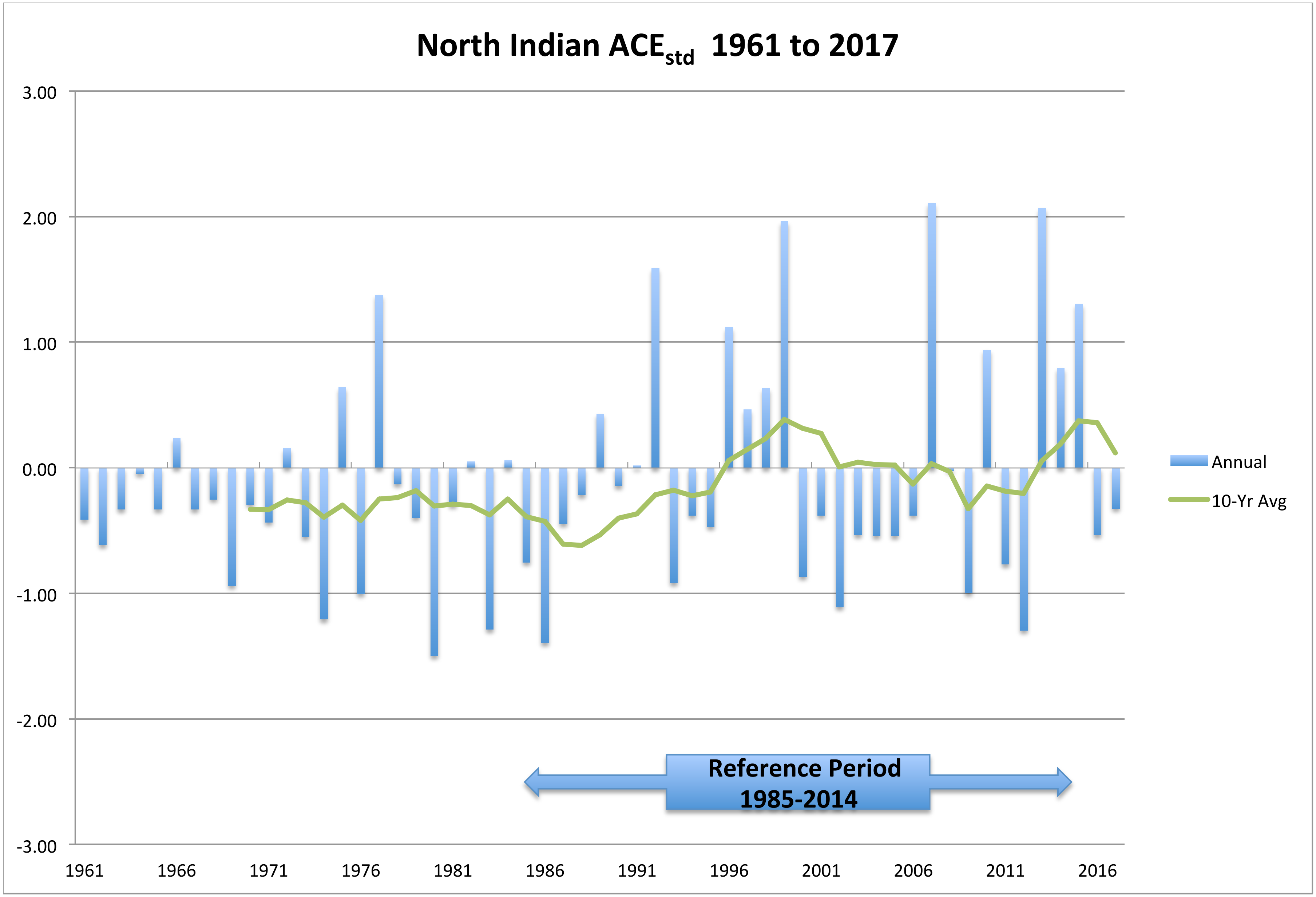


Exhibit 2
Sheet 6

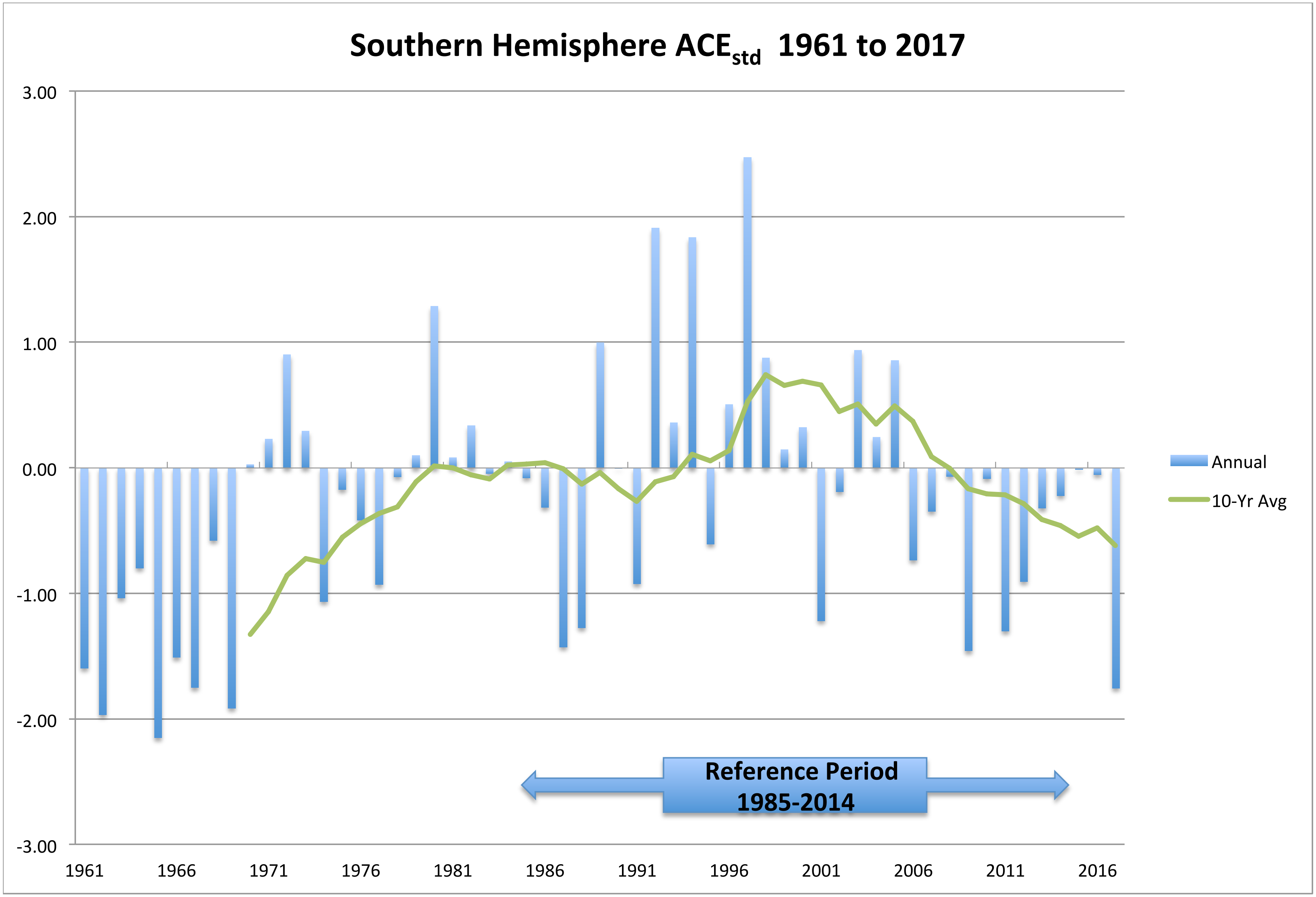


Exhibit 3
Sheet 1

Worldwide Accumulated Cyclone Energy as Standardized Anomaly (ACEstd)

	Reference Period 1961 - 1990				Reference Period 1985 - 2014			
	Worldwide ACEstd	5-Year Average	10-Year Average	20-Year Average	Worldwide ACEstd	5-Year Average	10-Year Average	20-Year Average
1961	0.38				-0.26			
1962	-0.82				-1.09			
1963	0.06				-0.48			
1964	0.81				0.04			
1965	-0.32	0.02			-0.74	-0.50		
1966	-0.28	-0.11			-0.71	-0.60		
1967	0.20	0.09			-0.38	-0.45		
1968	-0.20	0.04			-0.66	-0.49		
1969	-1.52	-0.43			-1.58	-0.81		
1970	-0.41	-0.44	-0.21		-0.80	-0.83	-0.67	
1971	1.46	-0.10	-0.10		0.49	-0.59	-0.59	
1972	1.52	0.17	0.13		0.54	-0.40	-0.43	
1973	-1.12	-0.01	0.01		-1.29	-0.53	-0.51	
1974	-1.50	-0.01	-0.22		-1.56	-0.52	-0.67	
1975	-0.90	-0.11	-0.27		-1.14	-0.59	-0.71	
1976	0.07	-0.38	-0.24		-0.47	-0.79	-0.69	
1977	-2.38	-1.16	-0.50		-2.17	-1.33	-0.86	
1978	0.40	-0.86	-0.44		-0.24	-1.12	-0.82	
1979	-0.20	-0.60	-0.30		-0.66	-0.94	-0.73	
1980	0.61	-0.30	-0.20	-0.21	-0.09	-0.73	-0.66	-0.66
1981	-0.44	-0.40	-0.39	-0.25	-0.83	-0.80	-0.79	-0.69
1982	0.99	0.27	-0.45	-0.16	0.16	-0.33	-0.83	-0.63
1983	-0.25	0.14	-0.36	-0.17	-0.69	-0.42	-0.77	-0.64
1984	0.86	0.35	-0.12	-0.17	0.08	-0.27	-0.60	-0.64
1985	0.36	0.30	0.00	-0.14	-0.27	-0.31	-0.52	-0.61
1986	-0.11	0.37	-0.01	-0.13	-0.59	-0.26	-0.53	-0.61
1987	-0.19	0.14	0.20	-0.15	-0.65	-0.42	-0.38	-0.62
1988	-0.76	0.03	0.09	-0.17	-1.04	-0.50	-0.46	-0.64
1989	1.40	0.14	0.25	-0.03	0.45	-0.42	-0.35	-0.54
1990	2.27	0.52	0.41	0.11	1.05	-0.16	-0.23	-0.45
1991	1.00	0.74	0.56	0.08	0.17	0.00	-0.13	-0.46
1992	4.42	1.67	0.90	0.23	2.54	0.64	0.11	-0.36
1993	0.56	1.93	0.98	0.31	-0.13	0.82	0.16	-0.30
1994	2.77	2.20	1.17	0.52	1.40	1.01	0.29	-0.16
1995	0.72	1.89	1.21	0.61	-0.02	0.79	0.32	-0.10
1996	1.92	2.08	1.41	0.70	0.81	0.92	0.46	-0.04
1997	4.12	2.02	1.84	1.02	2.34	0.88	0.76	0.19
1998	0.66	2.04	1.98	1.04	-0.06	0.89	0.85	0.20
1999	-0.37	1.41	1.80	1.03	-0.78	0.46	0.73	0.19
2000	0.19	1.30	1.60	1.01	-0.38	0.38	0.59	0.18
2001	-0.22	0.87	1.48	1.02	-0.67	0.09	0.50	0.19
2002	0.55	0.16	1.09	0.99	-0.14	-0.41	0.24	0.17
2003	1.41	0.31	1.17	1.08	0.46	-0.30	0.29	0.23
2004	2.84	0.95	1.18	1.18	1.45	0.14	0.30	0.30
2005	2.08	1.33	1.32	1.26	0.92	0.40	0.39	0.36
2006	0.49	1.47	1.17	1.29	-0.18	0.50	0.29	0.38

Exhibit 3
Sheet 2

Worldwide Accumulated Cyclone Energy as Standardized Anomaly (ACEstd)

	Reference Period 1961 - 1990				Reference Period 1985 - 2014			
	Worldwide ACEstd	5-Year Average	10-Year Average	20-Year Average	Worldwide ACEstd	5-Year Average	10-Year Average	20-Year Average
2007	-0.87	1.19	0.68	1.26	-1.12	0.31	-0.05	0.35
2008	-0.42	0.82	0.57	1.27	-0.81	0.05	-0.13	0.36
2009	-0.83	0.09	0.52	1.16	-1.09	-0.46	-0.16	0.29
2010	-0.93	-0.51	0.41	1.00	-1.16	-0.87	-0.24	0.18
2011	-0.90	-0.79	0.34	0.91	-1.14	-1.07	-0.28	0.11
2012	0.07	-0.60	0.29	0.69	-0.47	-0.94	-0.31	-0.04
2013	-0.51	-0.62	0.10	0.64	-0.87	-0.95	-0.45	-0.08
2014	0.75	-0.30	-0.11	0.54	0.00	-0.73	-0.59	-0.15
2015	3.33	0.55	0.02	0.67	1.79	-0.14	-0.51	-0.06
2016	0.95	0.92	0.06	0.62	0.14	0.12	-0.47	-0.09
2017	-0.73	0.76	0.08	0.38	-1.03	0.00	-0.47	-0.26

	1961-1990	1985-2014
Mean	677.3	765.6
Standard Dev.	117.8	169.9

ACEstd = (ACE for Year N - Reference Period Mean) / (Reference Period Standard deviation)

Exhibit 4
Sheet 1

Worldwide Accumulated Cyclone Energy as Standardized Anomaly (ACEstd) - By Region

	ACEstd - Reference Period 1985 - 2014									
	North Atlantic		East Pacific		West Pacific		North Indian		Southern Hemisphere	
	ACEstd	10-Year Average	ACEstd	10-Year Average	ACEstd	10-Year Average	ACEstd	10-Year Average	ACEstd	10-Year Average
1961	1.49		-1.60		0.55		-0.41		-1.59	
1962	-1.16		-1.58		1.04		-0.61		-1.97	
1963	0.13		-1.60		0.68		-0.33		-1.04	
1964	0.94		-1.69		0.90		-0.05		-0.80	
1965	-0.41		-1.46		1.15		-0.33		-2.15	
1966	0.55		-1.03		-0.05		0.24		-1.51	
1967	0.19		-0.98		0.86		-0.33		-1.75	
1968	-1.02		-1.03		0.49		-0.25		-0.58	
1969	0.88		-1.55		-0.95		-0.94		-1.91	
1970	-1.10	0.05	-0.81	-1.33	-0.15	0.45	-0.29	-0.33	0.03	-1.33
1971	-0.20	-0.12	0.20	-1.15	0.70	0.47	-0.43	-0.33	0.23	-1.14
1972	-1.16	-0.12	0.16	-0.98	0.91	0.45	0.16	-0.26	0.90	-0.86
1973	-0.97	-0.23	-0.21	-0.84	-1.45	0.24	-0.55	-0.28	0.29	-0.72
1974	-0.66	-0.39	-0.61	-0.73	-0.98	0.05	-1.20	-0.39	-1.06	-0.75
1975	-0.53	-0.40	-0.25	-0.61	-1.32	-0.19	0.64	-0.30	-0.17	-0.55
1976	-0.41	-0.50	-0.10	-0.52	-0.10	-0.20	-1.00	-0.42	-0.42	-0.44
1977	-1.30	-0.65	-1.76	-0.60	-1.30	-0.42	1.38	-0.25	-0.93	-0.36
1978	-0.75	-0.62	1.34	-0.36	-0.62	-0.53	-0.13	-0.24	-0.07	-0.31
1979	-0.26	-0.73	-1.17	-0.32	-0.24	-0.46	-0.39	-0.18	0.10	-0.11
1980	0.61	-0.56	-0.82	-0.32	-0.62	-0.50	-1.50	-0.30	1.29	0.02
1981	-0.16	-0.56	-0.92	-0.43	-0.72	-0.64	-0.30	-0.29	0.08	0.00
1982	-1.22	-0.57	0.57	-0.39	0.47	-0.69	0.05	-0.30	0.34	-0.05
1983	-1.46	-0.61	1.33	-0.24	-0.78	-0.62	-1.29	-0.37	-0.05	-0.09
1984	-0.41	-0.59	1.11	-0.07	-0.29	-0.55	0.06	-0.25	0.05	0.02
1985	-0.34	-0.57	1.10	0.07	-0.69	-0.49	-0.75	-0.39	-0.08	0.03
1986	-1.16	-0.65	-0.33	0.04	0.27	-0.45	-1.39	-0.43	-0.31	0.04
1987	-1.19	-0.63	0.11	0.23	0.49	-0.27	-0.44	-0.61	-1.43	-0.01
1988	-0.11	-0.57	-0.22	0.08	-0.71	-0.28	-0.22	-0.62	-1.27	-0.13
1989	0.39	-0.50	-0.26	0.17	0.00	-0.26	0.43	-0.53	1.00	-0.04
1990	-0.20	-0.59	2.05	0.45	0.66	-0.13	-0.14	-0.40	0.00	-0.17
1991	-1.16	-0.69	0.86	0.63	1.01	0.04	0.02	-0.37	-0.92	-0.27
1992	-0.53	-0.62	2.74	0.85	1.53	0.15	1.59	-0.21	1.91	-0.11
1993	-1.11	-0.58	1.25	0.84	-0.34	0.19	-0.91	-0.18	0.36	-0.07
1994	-1.22	-0.66	0.97	0.83	1.38	0.36	-0.38	-0.22	1.83	0.11
1995	1.84	-0.45	-0.46	0.67	-0.47	0.38	-0.47	-0.19	-0.61	0.06
1996	0.88	-0.24	-1.24	0.58	1.02	0.46	1.12	0.06	0.50	0.14
1997	-1.08	-0.23	0.74	0.64	2.44	0.65	0.46	0.15	2.47	0.53
1998	1.13	-0.11	0.12	0.68	-1.41	0.58	0.63	0.24	0.88	0.74
1999	1.05	-0.04	-0.62	0.64	-1.81	0.40	1.96	0.39	0.15	0.66
2000	0.14	-0.01	-0.53	0.38	-0.49	0.29	-0.86	0.32	0.32	0.69
2001	0.00	0.11	-0.62	0.24	0.02	0.19	-0.38	0.28	-1.22	0.66
2002	-0.67	0.10	-0.05	-0.04	0.44	0.08	-1.11	0.01	-0.19	0.45
2003	1.04	0.31	-1.19	-0.29	0.29	0.14	-0.53	0.04	0.94	0.51
2004	1.84	0.62	-0.94	-0.48	1.63	0.17	-0.54	0.03	0.25	0.35
2005	2.12	0.65	-0.52	-0.48	0.05	0.22	-0.54	0.02	0.86	0.50
2006	-0.42	0.52	0.50	-0.31	0.16	0.13	-0.38	-0.13	-0.73	0.37

Exhibit 4
Sheet 2

Worldwide Accumulated Cyclone Energy as Standardized Anomaly (ACEstd) - By Region

	ACEstd - Reference Period 1985 - 2014									
	North Atlantic		East Pacific		West Pacific		North Indian		Southern Hemisphere	
	ACEstd	10-Year Average	ACEstd	10-Year Average	ACEstd	10-Year Average	ACEstd	10-Year Average	ACEstd	10-Year Average
2007	-0.56	0.57	-1.26	-0.51	-0.78	-0.19	2.11	0.04	-0.35	0.09
2008	0.57	0.51	-0.73	-0.60	-1.16	-0.17	-0.02	-0.03	-0.07	0.00
2009	-0.89	0.32	-0.02	-0.53	-0.24	-0.01	-0.99	-0.33	-1.46	-0.16
2010	0.87	0.39	-1.28	-0.61	-1.69	-0.13	0.94	-0.14	-0.09	-0.21
2011	0.25	0.41	-0.11	-0.56	-1.06	-0.24	-0.77	-0.18	-1.30	-0.21
2012	0.36	0.52	-0.48	-0.60	-0.02	-0.28	-1.29	-0.20	-0.90	-0.29
2013	-1.16	0.30	-0.88	-0.57	-0.26	-0.34	2.07	0.06	-0.32	-0.41
2014	-0.67	0.05	1.28	-0.35	-0.26	-0.53	0.80	0.19	-0.22	-0.46
2015	-0.74	-0.24	2.70	-0.03	1.62	-0.37	1.31	0.38	-0.01	-0.54
2016	0.49	-0.15	0.97	0.02	-0.52	-0.44	-0.53	0.36	-0.05	-0.48
2017	1.82	0.09	-0.49	0.10	-1.39	-0.50	-0.32	0.12	-1.75	-0.62
1985-2014										
Mean ACE	109.9		127.3		304.1		20.1		204.2	
Standard Dev.	63.7		59.5		107.9		12.3		61.7	

ACEstd = (ACE for Year N - Reference Period Mean) / (Reference Period Standard deviation)

Exhibit 5
Sheet 1

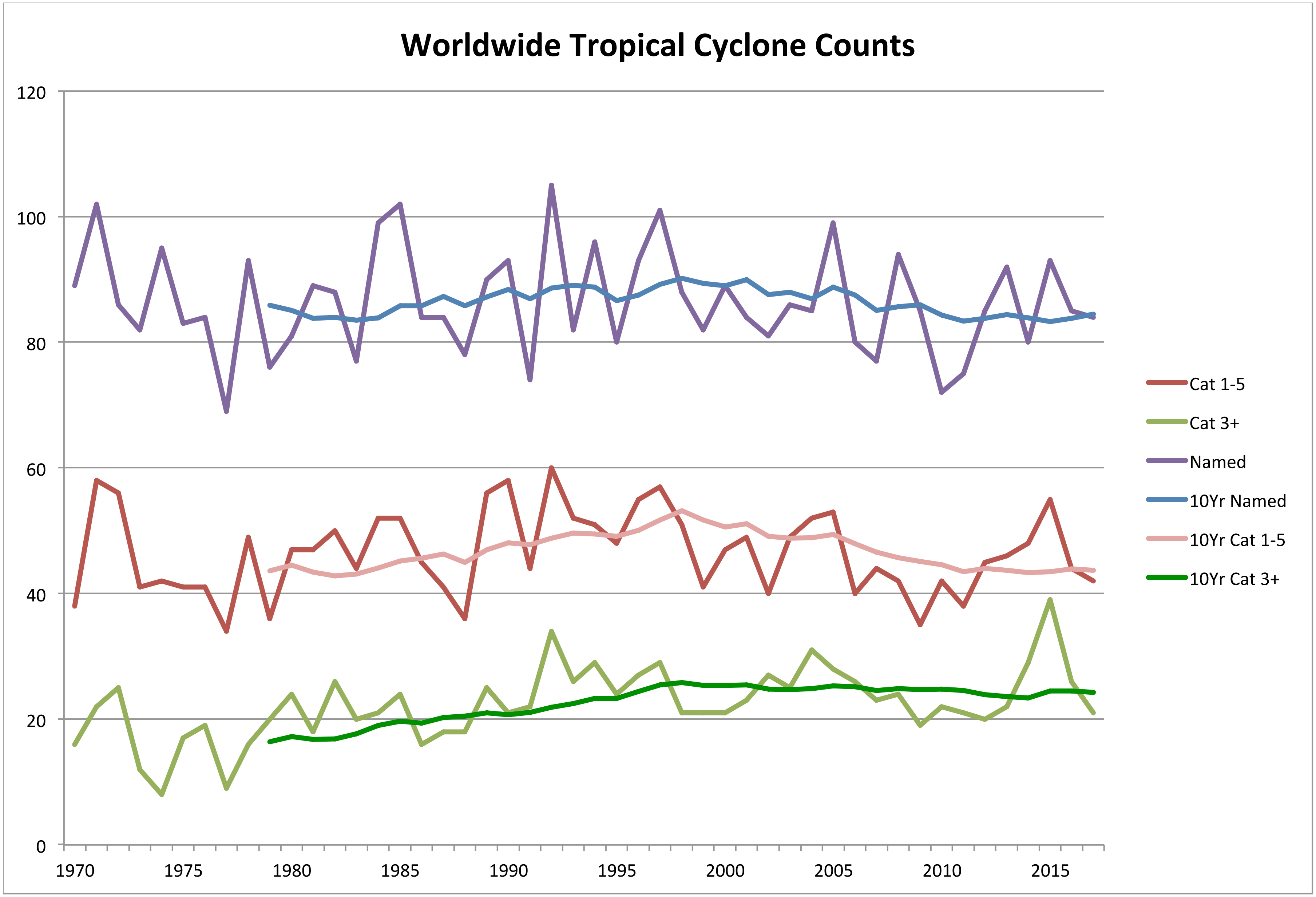


Exhibit 5
Sheet 2

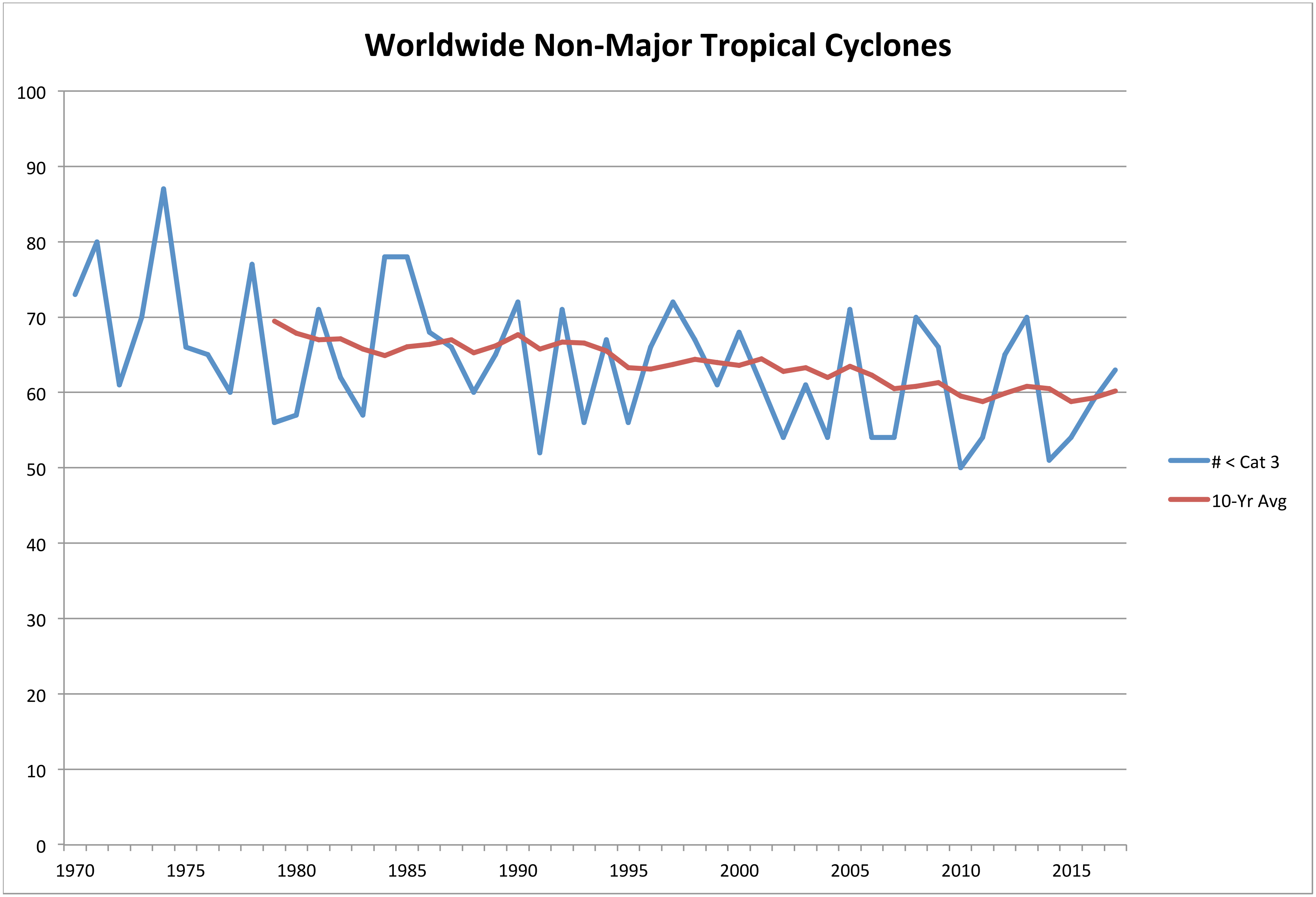


Exhibit 6
Sheet 1

Number of Tropical Cyclones

	Worldwide				Worldwide 10-Year Averages			
Year	Named Storms	Tropical Storms	Hurricanes, Cycl. & Typh.	Category 3+	Named Storms	Tropical Storms	Hurricanes, Cycl. & Typh.	Category 3+
1970	89	51	38	16				
1971	102	44	58	22				
1972	86	30	56	25				
1973	82	41	41	12				
1974	95	53	42	8				
1975	83	42	41	17				
1976	84	43	41	19				
1977	69	35	34	9				
1978	93	44	49	16				
1979	76	40	36	20	85.9	42.3	43.6	16.4
1980	81	34	47	24	85.1	40.6	44.5	17.2
1981	89	42	47	18	83.8	40.4	43.4	16.8
1982	88	38	50	26	84.0	41.2	42.8	16.9
1983	77	33	44	20	83.5	40.4	43.1	17.7
1984	99	47	52	21	83.9	39.8	44.1	19.0
1985	102	50	52	24	85.8	40.6	45.2	19.7
1986	84	39	45	16	85.8	40.2	45.6	19.4
1987	84	43	41	18	87.3	41.0	46.3	20.3
1988	78	42	36	18	85.8	40.8	45.0	20.5
1989	90	34	56	25	87.2	40.2	47.0	21.0
1990	93	35	58	21	88.4	40.3	48.1	20.7
1991	74	30	44	22	86.9	39.1	47.8	21.1
1992	105	45	60	34	88.6	39.8	48.8	21.9
1993	82	30	52	26	89.1	39.5	49.6	22.5
1994	96	45	51	29	88.8	39.3	49.5	23.3
1995	80	32	48	24	86.6	37.5	49.1	23.3
1996	93	38	55	27	87.5	37.4	50.1	24.4
1997	101	44	57	29	89.2	37.5	51.7	25.5
1998	88	37	51	21	90.2	37.0	53.2	25.8
1999	82	41	41	21	89.4	37.7	51.7	25.4
2000	89	42	47	21	89.0	38.4	50.6	25.4
2001	84	35	49	23	90.0	38.9	51.1	25.5
2002	81	41	40	27	87.6	38.5	49.1	24.8
2003	86	37	49	25	88.0	39.2	48.8	24.7
2004	85	33	52	31	86.9	38.0	48.9	24.9
2005	99	46	53	28	88.8	39.4	49.4	25.3
2006	80	40	40	26	87.5	39.6	47.9	25.2
2007	77	33	44	23	85.1	38.5	46.6	24.6
2008	94	52	42	24	85.7	40.0	45.7	24.9
2009	85	50	35	19	86.0	40.9	45.1	24.7
2010	72	30	42	22	84.3	39.7	44.6	24.8
2011	75	37	38	21	83.4	39.9	43.5	24.6
2012	85	40	45	20	83.8	39.8	44.0	23.9
2013	92	46	46	22	84.4	40.7	43.7	23.6
2014	80	32	48	29	83.9	40.6	43.3	23.4
2015	93	38	55	39	83.3	39.8	43.5	24.5
2016	85	41	44	26	83.8	39.9	43.9	24.5
2017	84	42	42	21	84.5	40.8	43.7	24.3
Total	4151	1917	2234	1075				
Average	86.5	39.9	46.5	22.4				
Total 1985-2014	2596	1179	1417	716				
Average	86.5	39.3	47.2	23.9				

Sources: http://tropical.atmos.colostate.edu/Realtime/
(except for East Pacific 1970 and North Indian 1970-1971 and South Atlantic: https://en.wikipedia.org/wiki/)

Exhibit 6
Sheet 2

Number of Tropical Cyclones

	North Atlantic				East Pacific			
	Named	Tropical		Category	Named	Tropical		Category
Year	Storms	Storms	Hurricanes	3+	Storms	Storms	Hurricanes	3+
1970	10	5	5	2	18	14	4	0
1971	13	7	6	1	18	6	12	6
1972	7	4	3	0	14	6	8	4
1973	8	4	4	1	12	5	7	3
1974	11	7	4	2	18	7	11	3
1975	9	3	6	3	17	8	9	4
1976	10	4	6	2	15	6	9	5
1977	7	2	5	1	8	4	4	0
1978	11	6	5	2	19	5	14	7
1979	9	3	6	2	10	4	6	4
1980	11	2	9	2	15	8	7	3
1981	12	5	7	3	15	7	8	1
1982	6	4	2	1	23	11	12	5
1983	4	1	3	1	21	9	12	8
1984	13	8	5	1	21	8	13	7
1985	11	4	7	3	24	11	13	8
1986	6	2	4	0	17	8	9	3
1987	7	4	3	1	20	10	10	4
1988	12	7	5	3	15	8	7	3
1989	11	4	7	2	18	9	9	4
1990	14	6	8	1	21	5	16	6
1991	8	4	4	2	14	4	10	5
1992	7	3	4	1	27	11	16	10
1993	8	4	4	1	15	4	11	9
1994	7	4	3	0	20	10	10	5
1995	19	8	11	5	10	3	7	3
1996	13	4	9	6	9	4	5	2
1997	8	5	3	1	19	10	9	7
1998	14	4	10	3	13	4	9	6
1999	12	4	8	5	9	3	6	2
2000	15	7	8	3	19	13	6	2
2001	15	6	9	4	16	8	8	2
2002	12	8	4	2	15	7	8	6
2003	16	9	7	3	16	9	7	0
2004	15	6	9	6	12	6	6	3
2005	28	13	15	7	15	8	7	2
2006	10	5	5	2	19	8	11	6
2007	15	9	6	2	11	7	4	1
2008	16	8	8	5	17	10	7	2
2009	9	6	3	2	20	12	8	5
2010	19	7	12	5	8	5	3	2
2011	19	12	7	4	11	1	10	6
2012	19	9	10	2	17	7	10	5
2013	14	12	2	0	20	11	9	1
2014	8	2	6	2	22	6	16	9
2015	11	7	4	2	26	10	16	11
2016	15	8	7	4	21	8	13	6
2017	17	7	10	6	18	9	9	4
Total	571	273	298	119	798	357	441	210
Average	11.9	5.7	6.2	2.5	16.6	7.4	9.2	4.4
Total 1985-2014	387	186	201	83	489	222	267	129
Average	12.9	6.2	6.7	2.8	16.3	7.4	8.9	4.3

Exhibit 6
Sheet 3

Number of Tropical Cyclones

	West Pacific				North Indian			
	Named	Tropical		Category	Named	Tropical		Category
Year	Storms	Storms	Typhoons	3+	Storms	Storms	Cyclones	3+
1970	24	12	12	11	7	4	3	1
1971	35	11	24	11	7	5	2	1
1972	30	8	22	14	4	0	4	0
1973	21	9	12	4	4	4	0	0
1974	32	17	15	3	1	0	1	0
1975	20	6	14	5	5	2	3	0
1976	25	11	14	9	5	5	0	0
1977	19	8	11	4	5	3	2	2
1978	28	13	15	3	4	2	2	0
1979	23	9	14	8	5	4	1	0
1980	24	9	15	9	3	3	0	0
1981	28	12	16	6	3	1	2	0
1982	25	6	19	12	5	3	2	1
1983	23	11	12	6	3	3	0	0
1984	27	11	16	9	4	2	2	0
1985	26	9	17	6	6	6	0	0
1986	27	8	19	8	3	3	0	0
1987	24	7	17	12	8	8	0	0
1988	26	13	13	7	5	4	1	1
1989	31	10	21	8	2	2	0	1
1990	32	11	21	8	2	1	1	1
1991	29	9	20	11	4	3	1	1
1992	32	11	21	11	10	7	3	1
1993	30	10	20	9	3	0	3	0
1994	36	16	20	12	5	4	1	1
1995	27	12	15	7	4	2	2	1
1996	36	15	21	10	8	4	4	1
1997	32	10	22	12	4	2	2	1
1998	18	9	9	5	8	3	5	1
1999	24	13	11	4	5	2	3	3
2000	25	10	15	8	4	2	2	0
2001	29	9	20	11	3	2	1	1
2002	24	8	16	11	5	5	0	0
2003	22	5	17	11	4	3	1	0
2004	31	11	20	14	5	3	2	0
2005	24	6	18	10	6	6	0	0
2006	22	9	13	10	6	5	1	1
2007	22	6	16	9	6	3	3	2
2008	27	15	12	8	6	5	1	1
2009	24	9	15	7	5	4	1	0
2010	15	6	9	4	5	2	3	2
2011	18	8	10	7	6	5	1	0
2012	25	10	15	10	4	4	0	0
2013	28	12	16	11	6	2	4	1
2014	21	9	12	7	5	3	2	2
2015	26	6	20	16	5	3	2	2
2016	24	11	13	11	5	4	1	0
2017	26	14	12	6	4	2	2	1
Total	1247	480	767	415	232	155	77	31
Average	26.0	10.0	16.0	8.6	4.8	3.2	1.6	0.6
Total 1985-2014	787	296	491	268	153	105	48	23
Average	26.2	9.9	16.4	8.9	5.1	3.5	1.6	0.8

Exhibit 6
Sheet 4

Number of Tropical Cyclones

	Southern Hemisphere				South Indian			
	Named	Tropical		Category	Named	Tropical		Category
Year	Storms	Storms	Cyclones	3+	Storms	Storms	Cyclones	3+
1970	30	16	14	2	17	6	11	2
1971	29	15	14	3	20	7	13	3
1972	31	12	19	7	15	7	8	2
1973	37	19	18	4	23	9	14	3
1974	33	22	11	0	18	10	8	0
1975	32	23	9	5	20	14	6	3
1976	29	17	12	3	16	9	7	3
1977	30	18	12	2	15	7	8	1
1978	31	18	13	4	20	11	9	4
1979	29	20	9	6	17	12	5	4
1980	28	12	16	10	16	4	12	9
1981	31	17	14	8	21	11	10	8
1982	29	14	15	7	19	10	9	2
1983	26	9	17	5	11	5	6	0
1984	34	18	16	4	24	13	11	3
1985	35	20	15	7	20	12	8	2
1986	31	18	13	5	21	12	9	5
1987	25	14	11	1	11	7	4	0
1988	20	10	10	4	13	7	6	2
1989	28	9	19	10	17	4	13	6
1990	24	12	12	5	17	8	9	4
1991	19	10	9	3	14	8	6	2
1992	29	13	16	11	14	7	7	7
1993	26	12	14	7	12	6	6	2
1994	28	11	17	11	22	9	13	7
1995	20	7	13	8	14	4	10	7
1996	27	11	16	8	18	5	13	7
1997	38	17	21	8	20	8	12	5
1998	35	17	18	6	15	9	6	2
1999	32	19	13	7	20	10	10	6
2000	26	10	16	8	17	5	12	7
2001	21	10	11	5	14	5	9	4
2002	25	13	12	8	17	7	10	7
2003	28	11	17	11	18	8	10	5
2004	22	7	15	8	15	3	12	6
2005	26	13	13	9	18	11	7	4
2006	23	13	10	7	15	10	5	5
2007	23	8	15	9	14	3	11	8
2008	28	14	14	8	21	11	10	5
2009	27	19	8	5	18	11	7	4
2010	25	10	15	9	14	6	8	5
2011	21	11	10	4	12	7	5	1
2012	20	10	10	3	16	7	9	2
2013	24	9	15	9	16	6	10	6
2014	24	12	12	9	14	5	9	7
2015	25	12	13	8	16	8	8	5
2016	20	10	10	5	9	4	5	3
2017	19	10	9	4	11	6	5	2
Total	1303	652	651	300	795	374	421	197
Average	27.1	13.6	13.6	6.3	16.6	7.8	8.8	4.1
Total 1985-2014	780	370	410	213	487	221	266	140
Average	26.0	12.3	13.7	7.1	16.2	7.4	8.9	4.7

Exhibit 6
Sheet 5

Number of Tropical Cyclones

	South Pacific				South Atlantic			
	Named	Tropical		Category	Named	Tropical		Category
Year	Storms	Storms	Cyclones	3+	Storms	Storms	Hurricanes	3+
1970	13	10	3	0	0	0	0	0
1971	9	8	1	0	0	0	0	0
1972	16	5	11	5	0	0	0	0
1973	14	10	4	1	0	0	0	0
1974	15	12	3	0	0	0	0	0
1975	12	9	3	2	0	0	0	0
1976	13	8	5	0	0	0	0	0
1977	15	11	4	1	0	0	0	0
1978	11	7	4	0	0	0	0	0
1979	12	8	4	2	0	0	0	0
1980	12	8	4	1	0	0	0	0
1981	10	6	4	0	0	0	0	0
1982	10	4	6	5	0	0	0	0
1983	15	4	11	5	0	0	0	0
1984	10	5	5	1	0	0	0	0
1985	15	8	7	5	0	0	0	0
1986	10	6	4	0	0	0	0	0
1987	14	7	7	1	0	0	0	0
1988	7	3	4	2	0	0	0	0
1989	11	5	6	4	0	0	0	0
1990	7	4	3	1	0	0	0	0
1991	4	1	3	1	1	1	0	0
1992	15	6	9	4	0	0	0	0
1993	14	6	8	5	0	0	0	0
1994	6	2	4	4	0	0	0	0
1995	6	3	3	1	0	0	0	0
1996	9	6	3	1	0	0	0	0
1997	18	9	9	3	0	0	0	0
1998	20	8	12	4	0	0	0	0
1999	12	9	3	1	0	0	0	0
2000	9	5	4	1	0	0	0	0
2001	7	5	2	1	0	0	0	0
2002	8	6	2	1	0	0	0	0
2003	10	3	7	6	0	0	0	0
2004	6	4	2	2	1	0	1	0
2005	8	2	6	5	0	0	0	0
2006	8	3	5	2	0	0	0	0
2007	9	5	4	1	0	0	0	0
2008	7	3	4	3	0	0	0	0
2009	9	8	1	1	0	0	0	0
2010	10	3	7	4	1	1	0	0
2011	9	4	5	3	0	0	0	0
2012	4	3	1	1	0	0	0	0
2013	8	3	5	3	0	0	0	0
2014	10	7	3	2	0	0	0	0
2015	9	4	5	3	0	0	0	0
2016	11	6	5	2	0	0	0	0
2017	8	4	4	2	0	0	0	0
Total	505	276	229	103	3	2	1	0
Average	10.5	5.8	4.8	2.1	0.1	0.0	0.0	0.0
Total 1985-2014	290	147	143	73	3	2	1	0
Average	9.7	4.9	4.8	2.4	0.1	0.1	0.0	0.0

A Simple Method for Modeling Changes Over Time

Uri Korn, FCAS

Abstract

Properly modeling changes over time is essential for forecasting and important for any model with data that spans multiple time periods. Regression models are probably the most commonly used for building predictive models in the insurance industry. These models do a fine job of fitting data and determining variable relationships, but are not meant for explaining how entities and relationships change over time, as time series models do. Time series models, on the other hand, have other drawbacks, depending on the type of model.

A method is presented to add time series components within a penalized regression framework so that these models are capable of handling everything a penalized generalized linear model can handle (distributional flexibility and credibility), as well as changes over time. Doing this, a subset of state space model functionality can be incorporated in a more familiar framework. The benefits of state space models in terms of their accuracy and intuitiveness are explained. This method can be useful for pricing models and detailed profitability studies, for example, as well as any other type of model with observations spanning multiple time periods.

Keywords. State Space Models, Penalized Regression, Elastic Net, Credibility, Forecasting, Time Series, Hierarchical Models

1. INTRODUCTION

Actuaries are frequently relied upon to make forecasts and predictions across time periods. This can be as a forecast for a future period, such as in ratemaking, as an interpretation of the past, as in reserving, or both together, as in profitability studies. Despite this, the most common type of model used, the linear regression model, is not equipped to handle changes over time, which is an important consideration when working with data that spans multiple time periods and when forecasting future periods.

This type of behavior is best modeled via a state space model, a flexible and powerful time series method. But besides being less familiar to many practitioners, they have other drawbacks as well, depending on how they are solved. A Bayesian model can be used, but these take extra time to build and fit and do not scale well to very large datasets. The other popular option is the Kalman Filter, but it is less accurate and does not provide the distributional flexibility that generalized linear models do.

This paper shows a method to add random walks and related state space model functionality to linear regression models. In a random walk, the complement of credibility for each period is the fitted value of the previous period. This is in contrast to including the time variable as a categorical variable, which would use the overall mean as the complement (if using a model that incorporates credibility, such as a mixed model). Besides being less intuitive since use as a categorical variable ignores the order of the periods, its performance is far inferior to the random walk, as is shown.

A simple method is shown that allows for a subset of state space model functionality, such as the described type of behavior, within a penalized linear regression framework that does not suffer the same drawbacks. This model can handle distributional flexibility and credibility, as well as time series components. With these modifications, these models are better equipped to deal with this type of data.

1.1 Research Context

State space models (SSMs) and penalized regression methods will be explained and used throughout this paper. On the SSM side, De Jong and Zehnwirth 1983 were the first to introduce their use into the actuarial literature and use them to smooth development patterns. Zehnwirth 1996 and Wuthrich and Merz 2008 both use SSMs to smooth reserving estimates and Evans and Schmid 2007 use them to smooth trend estimates. De Jong 2005 gives a nice overview of SSMs and also shows examples of their use in mortality modeling and claims reserving. Korn 2016 uses a simplified SSM to smooth loss ratios by year.

On the penalized regression side, see Hastie, et al. 2009. Hastie and Qian 2014 give a nice overview of these models and their use in the R modeling language. Williams et al. 2015 show the benefit of using these models for variable selection. And recently, Frees and Gee 2016 showed how these models can be used to price policy endorsements. These lists are not meant to be comprehensive; refer to the mentioned papers for further references.

1.2 Objective

The goal of this paper is to show an approach that fits within a regression framework, is capable of handling time series effects, works well with volatile data having relatively few periods, is capable of handling big data, and that produces results suitable for presentation. The proposed method will be referred to as a (linear) regression based state space model, or RSSM for short.

1.3 Outline

Section 2 gives an overview of SSMs. Section 3 discusses some alternative methods for handling changes over time and estimates their performance using simulation results. Section 4 explains how to implement a random walk using the RSSM. Section 5 discusses standardization of time series components, something that is necessary for penalized regression models. Sections 6 and 7 discuss more SSM functionality that can be implemented with this approach, including changing trends and momentum. Section 8 discusses some practical implementation issues, and section 9 demonstrates a use case involving yearly loss ratios.

2. AN OVERVIEW OF STATE SPACE MODELS

State space models (SSMs) are a commonly used methodology to model how different phenomena change over time. They are expressed as a series of related equations. Their flexibility and ease of interpretation make them a common modeling choice. They are usually solved for using either a Kalman Filter or a Bayesian type model (which are both discussed in the next section). The proposed approach provides another means of solving for a subset of SSMs within a GLM framework.

A simple linear trend model (known as "drift" in SSM terminology), for example, can be expressed as follows: (Kim and Nelson 1999)

$$Y_t = X_t + e_t$$

$$X_t = X_{t-1} + u$$

The first equation (also known as the measurement or observation equation) relates the actual data (*Y*) to the fitted values (*X*) with an error term (*e*). In the second (also known as the state or transition equation), the fitted values are increased by the trend (*u*) each period.

Another type of SSM is a random walk, which is a way to model gradual changes that can occur over time. The complement of credibility for each period is the fitted result of the previous period, which is an intuitive way to model changes over time. Such a model balances goodness of fit to the data versus having smaller or smoother changes from period to period. A random walk could be represented as follow:

$$Y_t = X_t + e_t$$

$$X_t = X_{t-1} + r_t$$

These equations are equivalent to the above except that in the second, the fitted values, instead of increasing by the same amount each period, are increased by varying amounts *(r)*. This variable is another error term whose values are also minimized. The result is a model that balances goodness of fit to the data with as little change as possible, depending on the ratio of the error terms. The first term *(e)* represents the volatility of the data, while the second *(r)* represents the variance or average magnitude of the period-to-period changes.

A model where the trend (or drift) itself changes via a random walk can be modeled as well. An example is as follows:

$$Y_t = X_t + e_t$$

$$X_t = X_{t-1} + u_t$$

$$u_t = u_{t-1} + r_t$$

Here, the third equation allows for the trend itself (*u*) to follow a random walk. Both *e* and *r* are error terms that are minimized.

More types of models are discussed as well. Even though the proposed models are based on state space models, they can still be used without a complete familiarity of SSMs.

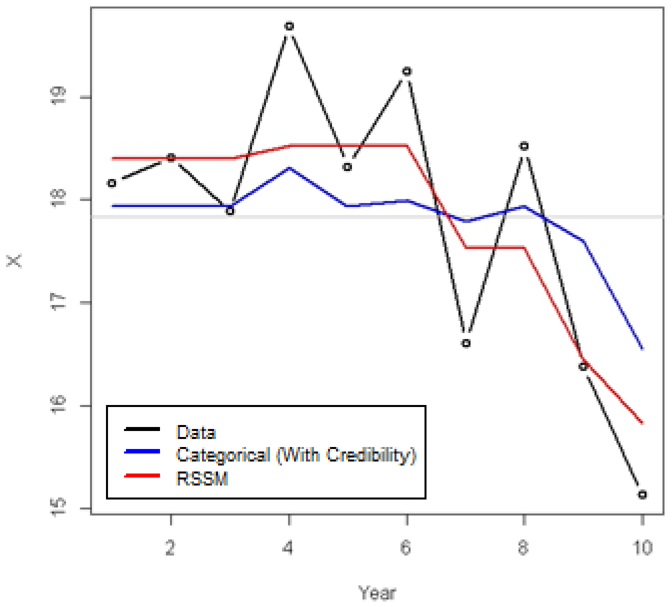
3. COMPARISON WITH EXISTING METHODS

3.1 An Overview

Perhaps the most common way that actuaries use to control for changes over time is to model the year within a GLM as a categorical variable. If a mixed model or penalized regression is used, credibility weighting is performed against the overall mean. But such an approach ignores the relationships between consecutive years. The complement of credibility for each year should be the fitted value of the previous year, which is much more intuitive than the overall mean. Figure 1 illustrates this point[1]. (A penalized regression model was used so that credibility is taken into account. This is similar to adding the year as a random effect in a mixed model.) Note the behavior in years 7 and 8, for example; even though the fitted curve should most likely be decreasing in this range, this does not occur since it is constrained to fall in between the data points and the overall mean. It can be seen that using an RSSM results in more intuitive behavior. Note how the former is also further off in the latest period making forecasts of future periods less accurate.

[1] All of these models were fit using an elastic net with 3-fold cross validation repeated 20 times.

Figure 1: Time as a categorical variable versus RSSM



State space models are another way to model changes over time. The most common methods of solving for an SSM are the Kalman Filter[2] and Bayesian Markov Chain Monte Carlo (MCMC) modeling. The Kalman Filter uses formulas to calculate the amount of credibility to be assigned to each period, using the previous period's prediction as the complement of credibility. The model requires three parameters, which are estimated via maximum likelihood: the value of the first period and two variance parameters that help determine the credibility. The calculations are made easier by assuming that both the distribution of the errors in the data as well as that of the period-to-period changes are normally distributed. (This model is essentially the time series equivalent of Buhlmann-Straub credibility.) For a more thorough review of the Kalman Filter, refer to Korn 2016.

One problem with using the Kalman Filter to model insurance data is its lack of distributional flexibility such as a GLM provides. Errors are assumed to be normally distributed and changes additive. There are some ways of fixing these issues (see Taylor and McGuire 2007 and Korn 2016), but these solutions are still not as robust, flexible, and/or as simple as the proposed. Using the Kalman Filter to fit the example data produces a fitted result with no changes, equal to the overall mean[3]. This is because this model requires more than just ten data points to adequately adapt to and fit the data.

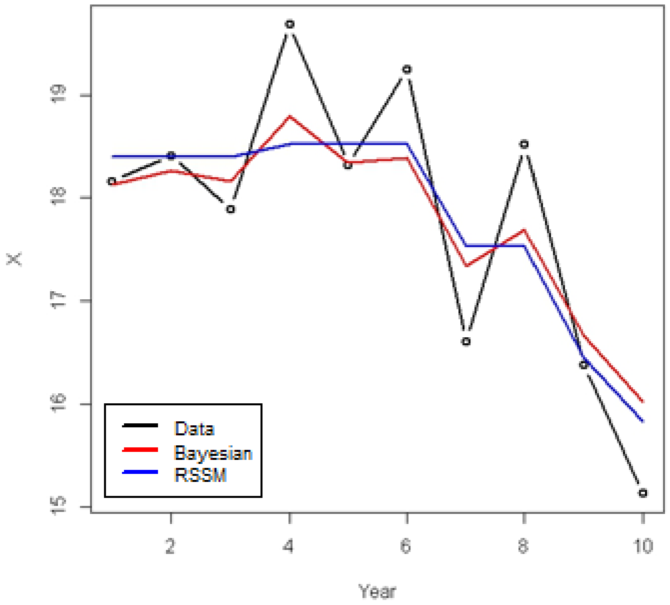
A more flexible framework is provided by Bayesian models, which are capable of modeling SSMs,

[2] The results of the Holt-Winters method, also known as exponential smoothing, should be roughly similar but less accurate than the Kalman Filter and will not be expanded upon here.

[3] The same was true when using the "bagging" method described in Korn 2016.

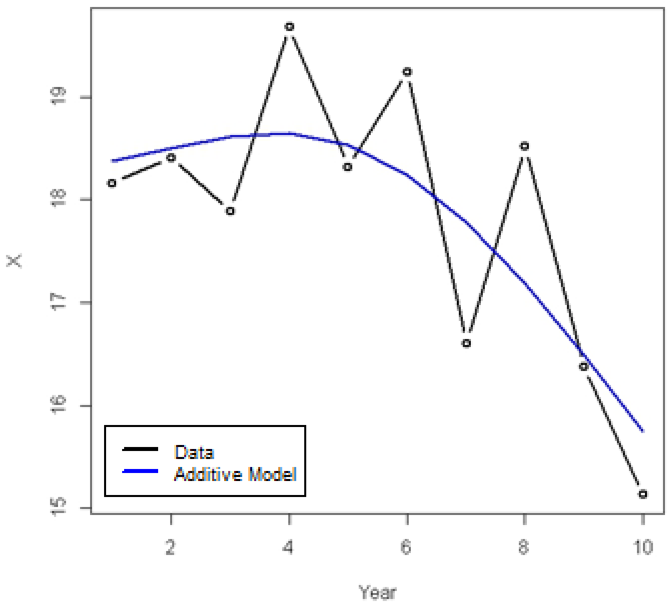
such as a random walk, as well as incorporating any type of distribution assumption. A Bayesian model implementing a random walk has parameters for every single period unlike the Kalman Filter that only has a parameter for the first period. Parameters are typically solved via MCMC techniques, which are simulations that are guided by the overall likelihood of the model. The downsides to these models are the specialized expertise required as well as the time needed to build and run each model. These models also do not scale well to large datasets or to a large number of parameters. As shown in Figure 2, running a Bayesian model on the example dataset performs satisfactory, although produces a much bumpier line than the proposed approach. This makes it more difficult to interpret and not as suitable for presentation. As can be seen, multiple changes are shown before year 6, despite little support for this in the data. The RSSM shows a decreasing trend starting from year 6, which seems to be the general trend of the data; the Bayesian line is still bumpy after this point.

Figure 2: Bayesian model versus RSSM



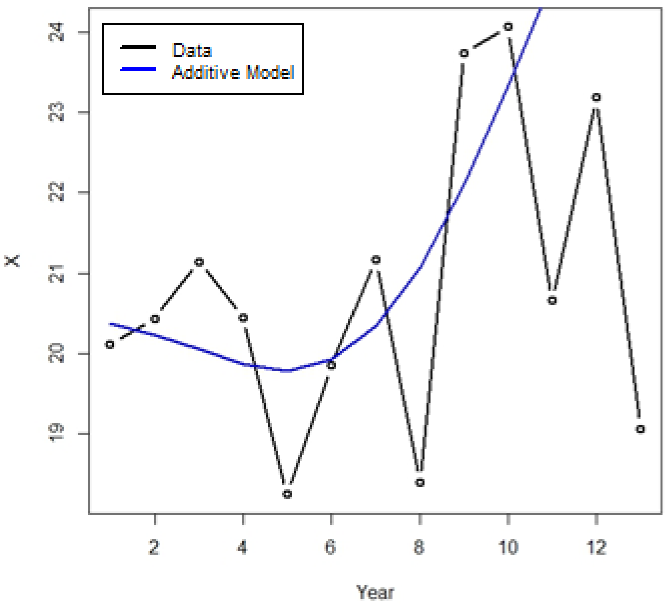
Another approach is to use an additive model, which uses a smoothing function, often a cubic spline, to adapt to the data. This type of model does a good job of fitting to the example data (using the mgcv package in R) as shown in Figure 3.

Figure 3: Additive model fit on the example data



A problem with splines is that even though the historical data may seem to fit well, they often show high trends at the end points, implying historical and prospective patterns that may not exist. Related to this, small changes in the data or a few new data points can often result in large changes. They are also very susceptible to outliers as shown in the next section. Figure 4 shows another example that demonstrates these issues. And finally, they are also difficult to blend with credibility techniques.

Figure 4: Issues with the additive model fit



Finally, there are ARIMA models. Besides for not being as intuitive as the methods discussed, they lack the flexibility of SSMs (Carlin 1992). Because of these issues, they will not be elaborated upon further.

3.2 Simulation Results

Simulations were conducted over a ten year period to compare the various methods to each other and to the proposed method. The first simulation exercise was fairly straightforward and did not attempt to mimic real data by including outliers, etc. The second simulation was meant to be more realistic and used t distributions instead of Gaussian and included occasional outliers to account for the fact that distribution fits are usually not exact. The results are shown in Table 1. The code used to run the simulations can be found in Appendix A.

Table 1: Simulation results of various time series methods

	Simulation 1 – No Outliers		**Simulation 2 – Outliers**		
Method	**RMSE[4] of Fitted Data**	**RMSE Relative to the Mean**	**RMSE of Fitted Data**	**RMSE Relative to the Mean**	**Difference in Percentages**
Mean	0.8414	0.0%	1.296	0.0%	
Elastic Net With Year as Factor	0.8118	-3.5%	1.412	+9.0%	+12.5%
Kalman Filter	0.8103	-3.7%	1.252	-3.3%	+0.4%
Kalman Filter with Bagging (See Korn 2016)	0.7660	-9.0%	1.196	-7.6%	+1.3%
Bayesian Model	0.6405	-23.9%	1.081	-16.6%	+7.3%
Additive Model[5]	0.6905	-17.9%	1.176	-9.2%	+8.7%
RSSM	0.6517	-22.5%	1.054	-18.7%	+3.9%
RSSM with 25% Momentum (See section 6.3)	0.6439	-23.5%	1.036	-20.0%	+3.4%

The Bayesian model was the best performing when there were no outliers, and the proposed method was the best with outliers, but both methods performed well. Note the poor performance of using the year as a factor (even with credibility being taken into account, as was the case here); when outliers are present, it is even worse than taking a simple average. The differences in the RMSE amounts are shown to the right as a rough measure of the robustness to outliers of each of the methods. Both the additive and Bayesian models are more susceptible to outliers than the proposed method. This is because these models depend on various formulas and assumptions to estimate the appropriate credibility, while the proposed uses cross validation and determines the best credibility by testing on the data itself.

[4] Root mean square error

[5] Note that this method is not completely compatible with the others, as it also has a trend component, which would improve the performance of the other models as well.

4. IMPLEMENTING A RANDOM WALK

4.1 Dummy Encodings

The proposed approach uses a GLM framework to implement a subset of SSM functionality, such as random walks. This provides a relatively simple and familiar modeling environment and also allows for distributional flexibility and credibility.

To explain the approach, when a categorical variable is added to a GLM, dummy encodings are created, such as those shown in Table 2. (The data values are shown on the left hand side, and the created model variables are shown across the top).

Table 2: Default dummy encodings for a year categorical variable

	2014	2015	2016
2013	0	0	0
2014	1	0	0
2015	0	1	0
2016	0	0	1

To implement a random walk, dummy encodings like those shown in Table 3 can be used instead.

Table 3: Initial dummy encodings for a random walk

	2014	2015	2016
2013	0	0	0
2014	1	0	0
2015	1	1	0
2016	1	1	1

With these, the coefficient value for 2014 affects not only that year, but the subsequent years, 2015 and 2016, as well. Likewise the coefficient value for 2015 effects both 2015 and the next year, 2016. If some form of credibility is applied (which is discussed in the next section), the starting point for each year is the previous year's fitted value. This allows for the fitted value of each year to be used as the complement of credibility for the following year. So, for example, if the 2015 coefficient is 0, its fitted value will match the 2014 fitted value.

Relating this back to SSMs, it can be seen that doing this is equivalent to the random walk, where *r* is the coefficient value for each year. The first equation (that relates the empirical data to the fitted values) is identical as well, except that here, the distribution of the error term (*e*) is determined by the GLM family.

$$Y_t = X_t + e_t$$

$$X_t = X_{t-1} + r_t$$

Implementing the approach in this fashion keeps the solution linear, which makes solving for the optimal parameters much easier. Non-linear problems are difficult to solve and even when solved, it is hard to determine if only a local maximum has been reached. Most statistical packages have methods for modifying the default dummy encodings of certain variables. Appendixes A and B show an example of doing this in R.

4.2 Penalized Regression and Cross Validation

Penalized regression will be used as the credibility technique for both the random walk and other coefficients. This works by imposing a penalty to the likelihood the more a coefficient deviates from zero, thus reducing the coefficient values. This pushes the fitted values back towards the intercept, which is the overall mean, and thus credibility is applied.

Unlike mixed models, for example, which use likelihood based formulas and a number of assumptions to determine the magnitude of the penalty parameters, penalized regression uses k-fold cross validation. K-fold cross validation works as follows: the data is randomly divided into *k* chunks, or "folds". The model is fit on *k - 1* of these folds using different values for the penalty parameter, and then each of these fitted models is tested on the remaining fold. This process is repeated *k* times, each time using a different fold for the validation. The penalty parameter that performs best on the test data is chosen. For smaller data sets, this process can even be repeated multiple times and the average penalty parameter selected. This procedure is implemented in many statistical packages.

As mentioned, this approach differs from Bayesian models, mixed models, and the Kalman Filter, all of which use different assumptions and formulas to estimate the penalty parameters. Another benefit of cross validation is that it provides an excellent framework for testing the results of the model as the cross validated predictions (that is, the predictions made on the holdout or test folds) can be compared against the actual data to calculate various metrics that are unaffected by any possible overfitting.

Another benefit is their ability to fit on a large number of variables, even with large datasets. This is because of the efficient fitting algorithm: the model is initially fit using a large penalty value that causes most of the coefficients to be near zero, making the model easy to solve for. This penalty is then gradually decreased and the model is refit, each time using the results of the previous model as the starting point for the coefficients. Because of this, changes to coefficient values are small at each step making it easier for the fitting procedure to find the optimal values (Friedman et al. 2009). Mixed models and Bayesian methods often do not scale as well with large datasets having a large number of parameters. The run time for mixed models, especially, deteriorates rapidly as the number of variables or data points increases.

4.3 Types of Penalized Regression Methods

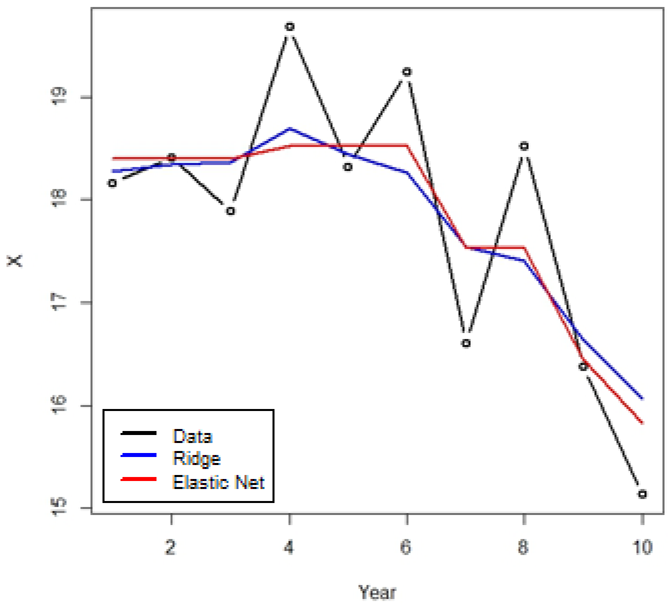
Penalized regression methods apply a penalty to the coefficient values in order to stabilize them. There are two types of penalty functions frequently used. Ridge regression is based on the squared value of the coefficients, also known as L2 regularization. This is similar to a mixed model or to using a normal distribution as the prior in a Bayesian setting.

The other type is the lasso, which penalizes based on the absolute value of the coefficients. A benefit of this type of model is that it can aid in variable selection. This is because the absolute value penalty will approach zero much faster than the squared values and so some coefficient values are set to zero, thus taking out their effect in the model. The downside of this model is that it does not handle correlated variables well. A compromise model called an elastic net provides the benefits of both by imposing a weighted average of both types of penalties (Zou and Hastie 2005). Such a model can handle correlated variables and perform variable selection.

There is another benefit of the elastic net for time series models. Because ridge regression does not shrink its coefficients down to zero, when using it to model a random walk, it will often show small changes in each year, even if it seems that there have not been any real changes. But the lasso cannot be used, since time series variables have correlations, thus making the elastic net the ideal choice[6]. A comparison of using ridge regression and the elastic net is shown in Figure 5. Note how ridge regression produces the bumpier line.

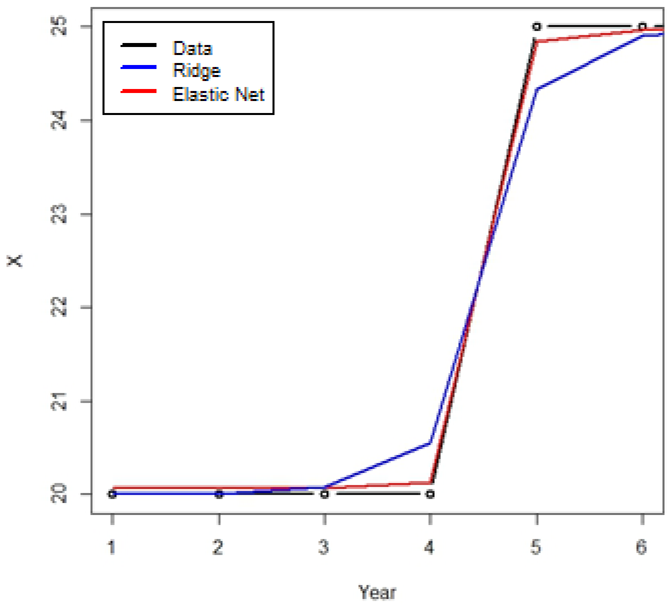
[6] Because of the first reason, it is suggested to give more weight to the lasso penalty.

Figure 5: Elastic net and ridge regression comparison



Related to this, because the square of a large number is an even larger number, if the data has a large jump in a single year, ridge regression will often model smaller changes over the course of several years. An elastic net model typically handles this scenario much better, as shown in Figure 5. The ridge model shows the decreasing trend as starting from year 4, where in reality, it seems that the decreasing trend did not start until year 6. A starker example is shown in Figure 6. It can be seen that a one-time increase is modeled over a period of three years.

Figure 6: Ridge regression comparison of a large change



For these reasons, the elastic net is the recommended model to use when fitting RSSMs.

4.4 Multiple Segmentations

Using the modified dummy encodings shown to model a random walk within a GLM framework makes it easy to not only model the overall changes, but also the changes by various segmentations. This can be done by including an interaction term between the segmentation and the random walk variable. Including a random walk variable by itself as well as an interaction term between the segmentation and the random walk produces a model that credibility weights each segment's changes using the overall average changes as the complement. This can be a powerful tool for handling yearly or quarterly data in a hierarchical fashion, much more detailed than simply modeling on an average trend.

A problem exists, however, when using the simple random walk dummy encoding shown above (Table 3) to model multiple segmentations. The issue is demonstrated in Figures 7 and 8. In this example, several segments are fit with their changes modeled using a random walk (using an interaction term as described) as well as a term for each segment.

Figure 7: A segment moving away from the mean

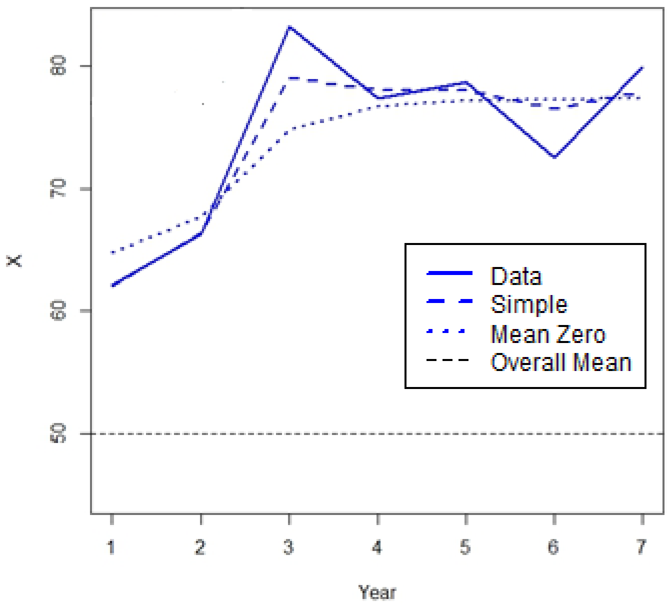
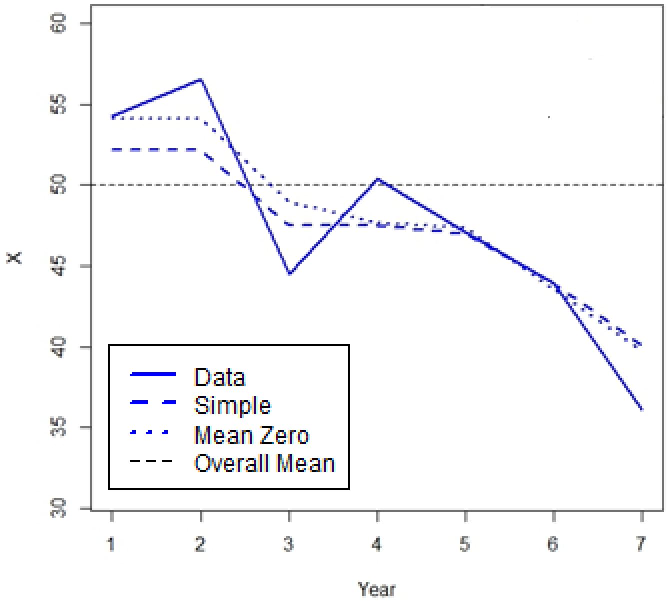


Figure 8: A segment moving towards the mean



The segment shown in Figure 7 is moving away from the overall mean, and in Figure 8, is moving towards the mean, as can be seen. Note how the fits using the simple encodings gives rise to a steeper curve than expected when moving away from the mean, and a flatter than expected curve when moving towards the mean. This is because when using the simple encodings, the first point of each segment is determined by the value of the segment coefficient (since all of the random walk variables are zero at this point). The subsequent points are determined from the interaction between the segment and the random walk variable, all of which are shrunken towards zero due to the penalty. Because of this, the model will often "take a shortcut" and gradually approach the data points over several periods in order to reduce the total value of this penalty, which results in the pattern mentioned.

To fix this issue, the random walk dummy encodings can be modified so that the mean of each year is zero (which can be done by simply subtracting the mean from each column). If these encodings are used instead, the segment coefficients now represent the average value of each segment, since the net effect of the random walk parameters is zero. Doing this fixes the tendency to "take a shortcut" and results in behavior that is more intuitive, as can be seen

On another note, it is worth mentioning that both the random walk and the segment parameters share the same penalty value. This does not mean that they will receive the same amount of credibility since this depends on other factors as well. But still, this should not be a concern as it is consistent with the penalized regression methodology, which uses the same penalty value for all parameters. If the variables are on the same scale (which they should be – see section 5), this will give them equal treatment in the model. An equal amount of explanation is penalized the same amount regardless of which variable it comes from. However, if desired, a different penalty can still be used for the random walk components by setting the penalty of these variables to a factor of the overall penalty and using another round of cross validation to determine this factor, although it is usually not necessary to do so.

4.5 Using Cross Validation with Panel Data

Panel data is the term used to describe data that both uses explanatory variables and has multiple observations across time periods, such as the data being described here. One of the assumptions of cross validation is that the folds are not correlated with each other. But this may be violated for this type of data, since the same entity may exist in different folds at different time periods, and these are correlated with each other.

To address this issue, instead of randomly selecting individual rows for inclusion in each fold as normally done, the entities themselves can be randomly assigned to folds along with all of their corresponding rows. This will reduce the correlation across folds when using panel data with cross validation. (Note that this is only necessary when a corresponding variable is not included in the model.)

4.6 Numeric Variables

This section illustrated how a random walk can be used to allow the coefficients of categorical variables to change over time. It is possible to do the same with numerical variables as well. Instead of including interactions, as done above, a new variable can be created and added to the model that is the product of the random walk and that variable. This works since:

$$\text{Coef1} \text{ x } \text{V} + \text{Coef2} \text{ x } \text{V} \text{ x } \text{RW} = \text{V} \text{ x } (\text{ Coef1} + \text{Coef2} \text{ x } \text{RW})$$

Where *V* is the desired variable, *RW* is a random walk variable, and *Coef1* and *Coef2* are two model coefficients. This allows for modeling how the relationships of numerical variables can change over time.

5. STANDARDIZATION

Before delving into more types of state space models, it is first necessary to discuss the standardization of different time series components. Since the same penalty is applied to every model variable, they should be on the same scale so that they receive equal treatment. Otherwise, equal coefficient values will cause greater changes to the variable with higher values. If one variable is stated in pounds and another in ounces, for example, the one stated in pounds has a larger scale and is likely to have greater effect on the fitted results. The most common approach is to normalize each variable by subtracting the mean and dividing by the standard deviation. This applies to numerical variables only. When dealing with both numerical and categorical variables, Gelman 2008 suggests dividing each numerical variable by twice the standard deviation instead. This is because a binary variable with fifty percent ones has a standard deviation of 0.5.

None of these approaches are designed for handling time series variables, however. These also need to be adjusted so that they can receive equal treatment. The following rules will be used to standardize time series variables to put them on the same scale as the other variables in the model and as each other:

1. A random walk variable (with no momentum, which is discussed later on) does not need to be adjusted, since it is similar to a categorical variable, which will not be modified. Since this random walk variable is not adjusted, all other time series variables can be compared to it for calculating their relative scaling factor.
2. Instead of comparing the standard deviations from the mean, the scale of a time series variable should be determined by calculating the square root of the average squared

differences between each time period. This is similar in nature to the common practice of using the standard deviation, but more properly reflects the nature of these variables.

When calculating these averages, since the denominator is the same for all time series variables (equal to the number of time periods), it will cancel out when being compared and can be ignored. This means that instead of comparing the average squared differences, the sum of the squared differences will be used instead. This quantity is equal to one for a plain random walk (with no momentum), which is the point of comparison. Therefore, the standardization divisor for each variable is simply equal to the square root of the sum of the squared differences[7].

Applying these rules to a simple trend (or drift) variable, which is a numerical sequence from one to the number of time periods, this variable would be divided by the square root of one less than the number of time periods. So, for example, a ten year series would be divided by three and a twenty six year series would be divided by five. (Note that the longer series receives a greater divisor. To explain, if both a trend and a random walk variable are in a model, the total penalty for using the random walk equals: $(n – 1)$ times the average change, where n is the number of time periods (assuming a lasso penalty for simplicity). The penalty for using the standardized trend variable equals: $\sqrt{}(n – 1)$ times the selected trend. Using the trend instead of the random walk can result in a lower penalty, but is also less flexible than the random walk. So, since the total penalty for using the random walk grows with the length of the sequence, to put the variables on equal footing, it is necessary for the trend penalty to do the same. Also, as the number of data points grows, a trend parameter is capable of having a greater impact on the likelihood, and so can withstand a larger penalty value.)

Both Appendix A and Appendix B show R code that use these standardization rules. When using in a penalized regression model, it is recommended to manually standardize all variables as described and to make sure that the penalized regression function used does not apply any additional standardization by default.

To recap, categorical variables should not be adjusted, numerical variables should be divided by twice the standard deviation, and time series variables should be divided by the square root of the sum of the squared differences.

[7] Note that twice this amount is not used, similar to how numerical variables are divided by twice their standard deviation, since dividing by this divisor already puts the variables on the same scale as a plain random walk. By contrast, it is necessary to divide numerical variables by twice their standard deviation to put them on the same scale as a categorical variable.

6. EXTENDING THE RANDOM WALK

6.1 Random Walk with Drift

The random walk model discussed above in section 4 assumes that the expected change for each period is zero, and this serves as the complement of credibility. If not the case, a trend (or drift) term can be added, and this value will serve as the effective complement of credibility instead. Such a term can be added to a GLM by including the time period as a numerical variable. (It is a good idea to set the mean of this variable to zero for the reasons mentioned in section 4.4. This variable should also be standardized as illustrated in section 5.)

6.2 Modeling a Changing Trend

All of the discussion above focused on a random walk on the level of a variable. It is also possible to model a changing trend (or drift) by using dummy encodings like those shown in Table 4.

Table 4: Example dummy encodings for a random walk on the slope

	2014	2015	2016
2013	0	0	0
2014	1	0	0
2015	2	1	0
2016	3	2	1

Using these, the 2014 coefficient, for example, will cause increases in years 2014 to 2016, and the 2015 coefficient will cause increases in years 2015 and 2016. This will cause a change to the slope. A trend term should also be added for the starting slope unless it is assumed to be zero. (Once again, coefficients that sum to zero should be used. The variable should also be standardized, as illustrated in section 5.)

6.3 Mean Reversion and Momentum

It is also possible to build a model that uses the concept of mean reversion. Allowing for mean reversion on the trend, for example, allows the trend to change, but also causes any changes to gradually decay over time and revert back to the long term average trend. This can be used to model shorter term changes in the trend that gradually revert back towards a long term average value. An example of dummy encodings with 25% mean reversion is shown in Table 5:

Table 5: Example dummy encodings with 25% mean reversion

	2014	2015	2016
2013	0	0	0
2014	1	0	0
2015	$1 + 0.75$	1	0
2016	$1 + 0.75 + 0.75^2$	$1 + 0.75$	1

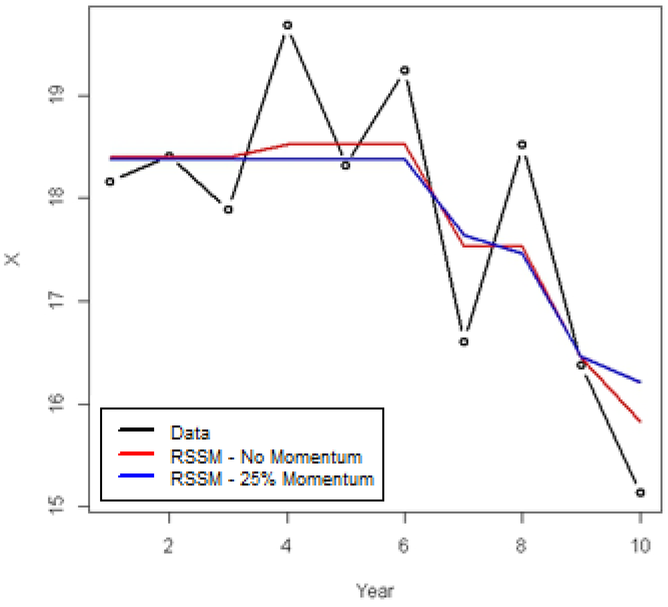
As can be seen, instead of adding one to each subsequent year, the added amounts decay exponentially. (As mentioned, after these dummy encoding are created, the mean should be subtracted from each column so that they all equal zero. They should be standardized as well as discussed in section 5. Those rules can be applied to standardize each column, which will receive slightly different penalties each.)

This concept of mean reversion can be used to relate a random walk on the level of a variable to a random walk on the trend. If the mean reversion is set to zero, no mean reversion will occur and the result is identical to a random walk on the trend. Alternatively, if the mean reversion is set to one, the changing trend will immediately revert back to its long term value after one period and so each change only affects a single period. This is identical to a random walk on the level of a variable. Any value in between zero and one can be viewed as a compromise of the two models.

One way of looking at these models (with perhaps a higher mean reversion value, although not necessarily) is as a random walk on the variable's level, but with momentum. In these models, the complement of credibility for the change of each period is a value between zero and the previous period's change. This will cause changes to continue in the same direction the following period, unless they are reversed. This is often a more realistic expectation since, quite often, changes display serial correlation over time.

Fitting the example data with this type of model produces the result shown in Figure 9. Note how this model both improves the fit to the data and results in a smoother, nicer looking curve.

Figure 9: Random walk with momentum



Cross validation can be used for choosing the optimal momentum for a model. Values can be tested in jumps of 5%, 10%, or 25%, etc., with all random walk variables sharing the same value. Or alternatively, multiple random walk variables with different momentum values can all be included and a grouped lasso penalty (which will cause each group of variables to either all be included or all be excluded) can be used to decide which are optimal, if supported by the statistical package being used.

The SSM equations for this mean reversion model are as follows (where the average long term trend is assumed to be zero for simplicity). It is easy to verify that these will produce identical results as using the dummy encodings shown above.

$$Y_t = X_t + e_t$$

$$X_t = X_{t-1} + u_t$$

$$u_t = au_{t-1} + r_t$$

It can be seen that if the mean reversion parameter (*a*) is set to one, these equations will be equivalent to those of a changing slope. If *a* is set to zero, then, $u_t = r_t$ and these equations are equivalent to a random walk on the level. If *a* is set to a value between zero and one, the trend will gradually decay back towards zero (or to the long term trend, if specified in the model).

6.4 Level Mean Reversion

It is worth mentioning that mean reversion can be used on the level of a variable as well, not just on the trend. This would cause any changes in the random walk to gradually decay, causing the level to revert back towards its long term average over time. If a trend or drift component is also included, the level will gradually revert back towards the trended long term average. This would be done by having the encodings of the random walk start at one as usual, but subsequent values are multiplied by a factor causing them to decay exponentially back towards zero over time. However, level mean reversion probably has less applicability to insurance modeling.

6.5 Extra Dispersion

It is also worth mentioning that another time series component can be added to provide some extra flexibility. A random walk models changes by period that are expected to continue in the next. In contrast, another component can be added for spikes and dips to the fitted values that occur only within a single period and which do not continue to the next. The SSM equations for this model are as follows:

$$Y_t = X_t + e_t$$

$$X_t = Z_t + d_t$$

$$Z_t = Z_{t-1} + r_t$$

Where Z_t is another state variable and d_t is this new component, which is an error term that is minimized. This component can be added to a GLM by including the year as a factor. However, doing so in addition to a random walk variable usually results in poorer performance (in the author's experience) and using it is not recommended, unless perhaps a greater penalty is applied to these variables. Another way to view this component is as a random walk with full level mean reversion.

7. MORE SSM COMPONENTS

7.1 Seasonality

If modeling on a period of less than a year, it may be necessary to account for the different levels of each month, quarter, or other unit, depending on the data. This can be accomplished simply by adding another categorical variable, adding another random walk, or by using splines[8].

7.2 Predictive Variables

Sometimes, the causes of yearly changes are understood and can be related to external variables. When this is the case, the variable can be incorporated in the model to help improve the predictions. This variable should be included as an index so that only changes in the variable affect the level each period and not the actual value of the variable (although if this variable is the same for every segment, the effect is identical).

To give an example, if yearly loss ratios by country are correlated with the interest rate, an index based on the interest rate can be used. This index can be created by dividing all values by the interest rate of the first year for that country. Note that the index was based on the interest rate itself and not the change in the interest rate so that changes in the interest rate will cause changes to the yearly loss ratios. If the interest rate has a lagged effect on the loss ratios, it is also possible to insert it lagged by the appropriate number of periods, which can be determined via another round of cross validation.

If, on the other hand, changes in the loss ratio trend are correlated with changes in the interest rate, for example, another variable can be added that is the product of the interest rate index and a numerical variable for the year (in addition to the year variable by itself to represent the average slope). This will work since:

$$\text{Coef1 x Y} + \text{Coef2 x I x Y} = \text{Y (Coef1} + \text{Coef2 x I)}$$

Where *Y* is the year, *I* is the interest rate index, and *Coef1* and *Coef2* are two model coefficients. As can be seen, changes in the index will cause changes in the slope, the amount of which is determined by the value of *Coef2*. (Both the interest rate and the slope should be standardized, as discussed in section 5.)

[8] Note that the issues mentioned above regarding additive models do not apply here since periods are repeated multiple times.

7.3 Multidimensional Random Walks

A two (or more) dimensional random walk can be constructed as well by interacting two random walks with each other. Depending on the packages used, the columns may need to be constructed manually, however. (The columns should be multiplied together while still in ones and zeros, and they can be made to sum to zero afterwards.) This can be useful for geographical smoothing, for example.

7.4 Correlated time series

Correlated time series can also be modeled using this framework. As an example, consider the case discussed above (section 7.2) where changes in the interest rate affect the loss ratio. As an alternative, the interest rate and the loss ratio can be modeled as correlated time series. This can be done by, instead of including the interest rate as a variable in each row, the interest rates would be given their own rows. Then, a model can be fit with a random walk that affects both the loss ratios and the interest rates. If the random walk variables in the loss ratio rows are multiplied by 10%, for example, each modeled point change in the interest rate will cause a 10% change in the loss ratio. The difference between this model and the one discussed above is that what the model determines to be the "errors" in the interest rate will not affect the loss ratio, since these do not affect the path of the random walk. It is also possible to assign a separate random walk to the loss ratios to capture the uncorrelated changes. (Although, in this example, different distribution families may be needed for each component, which is not possible with most standard regression packages. A categorical variable is also needed to determine whether each row is a loss ratio or an interest rate so that each can receive proper treatment in the model formula.)

Another example of a correlated time series approach is a dynamic factors model, which is a method for modeling missing or unknown variables that change over time (Geweke 1977). The missing variables are modeled via a random walk but coefficients by entity control the magnitude of the change for each entity. So essentially, this dynamic factor is a random walk whose magnitude varies by entity. An example of this is the stock market beta for each industry and company. Various market effects drive the value of stocks up and down, but each industry and company is affected differently from these changes.

For the simple interest rate model, the correlation percentage can be determined via another round of cross validation. Or alternatively, it can be approximated by initially fitting the percentage using the interest rate as a predictive variable, as discussed, then fitting another model for the random walk using this percentage, and then refitting the percentage using the fitted (standardized) random walk as a variable in the model.

A (mostly) one-sided dynamic factors model (where all coefficients are assumed to have the same sign) can be approximated in a similar fashion. A two-sided model can be approximated by initially fitting the random walk with a small penalty value so that it is not shrunken to zero. The penalty can then be gradually decreased as iterations between the random walk and the correlations are fit.

For more complicated models and for more exact solutions to these models, the EM algorithm (Expectation-Maximization, Dempster et al. 1977) can be used. Further discussion is outside the scope of this paper.

8. SOME NOTES ON IMPLEMENTATION

The method presented can be implemented using most existing elastic net packages. To do so, the default dummy encodings for random walk variables can be changed, as discussed. (See Appendix A and Appendix B for example R code.) If not possible, the random walk variables for each period can also be created and added manually.

Before a GLM solving algorithm is run, a matrix is created for the specified independent variables. Many GLM functions do this implicitly. Just as a separate column is created in this matrix for each possible value of a categorical variable, a separate column is also created for each period in a random walk variable, except for the first. If interactions with some segmentation are included, separate columns will be created for every combination of year and segment. Because of this, the resulting matrix can become quite large for models using a large number of predictors and that have a large number of data points. For most models, this is not an issue, but for very large ones, if encountering memory issues, instead of creating a standard matrix that utilizes memory for each cell value, a sparse matrix can be created instead, which only utilizes memory for non-zero cells. This can reduce the amount of memory required dramatically since most of the values in a typical modeling matrix are usually zero. (The example shown in Appendix A takes this approach.)

Some sparse matrix implementations (such as the "sparse.model.matrix" function in the "Matrix" package in R), when building a sparse matrix, will initially create a non-sparse matrix and only convert it to a sparse matrix at the very end. This can still create memory issues for very large models. If encountering issues, the matrix can be constructed manually one variable at a time so that sparse matrices can be used even during construction, which will save memory.

9. LOSS RATIO CASE STUDY

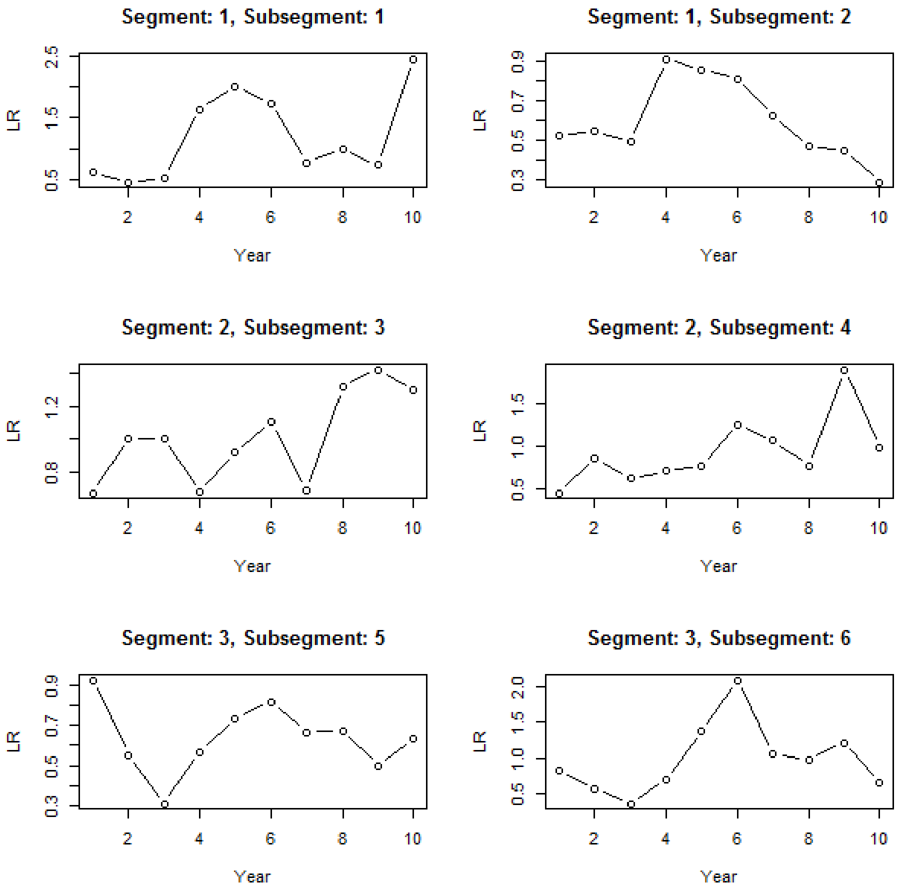
The proposed method will be demonstrated on an example involving yearly loss ratios. Three segments are used in the example, each having two sub-segments, making six sub-segments in all.

Each of these segments and sub-segments are affected by various changing factors, both on the premium and on the loss side, which cause the loss ratios to vary by year. Premium is affected by rate changes, which are easily accounted for, but the recorded rate changes are usually not a hundred percent accurate and do not take everything into account, such as changes in policy wording. On the losses side, besides from a longer term trend that affects the losses by a (perhaps) similar amount each year, social, legal, economic, and other factors can cause changes over shorter periods. Some of these may affect the entire book while others can be limited to various segments or sub-segments.

Another consideration is that claims take time to be reported and settled and so our current snapshot of losses will develop over time. Our goal is to estimate the ultimate loss ratio for each year and for future years for reserving, rate making, profit study, or informational purposes.

The ultimate chain ladder loss ratios are shown in Figure 10. This example assumes that premiums have already been on-leveled for rate changes and that losses have been trended by the long term average trend (although it is possible to use the procedure to fit this as well). For simplicity, the on-level premiums for each sub-segment for each year are assumed to be $1,000, although varying premiums can be accommodated as well.

Figure 10: Trended, ultimate loss ratios by segment



To fit this data, an elastic net with variables for a random walk is used. The model will account for the different loss ratio levels for each segment and sub-segment as well as the changes to each by year, incorporating credibility for both the level and changes. A Tweedie family is used to fit the aggregated losses.

A Cape Cod-like approach is used to account for development where the loss ratios inputted into the model are the reported loss ratios multiplied by the loss development factors (i.e., the chain ladder loss ratios), and the premiums divided by the loss development factors (i.e., the "used premiums") are used as the regression weights. This procedure accounts for development while taking into account the extra volatility of the greener years. This causes the model to give less weight to these years, but all years are still taken into account for determining the fitted loss ratios for each year. (As noted in Korn 2016, if full credibility is given to the yearly changes, the final indications will equal the chain ladder. If no credibility is given to the yearly changes, the results will equal the weighted average, which is the Bornhuetter-Ferguson loss ratio with the Cape Cod loss ratios used as the a priori. Anywhere in between can be thought of as a credibility weighting between these two methods.)

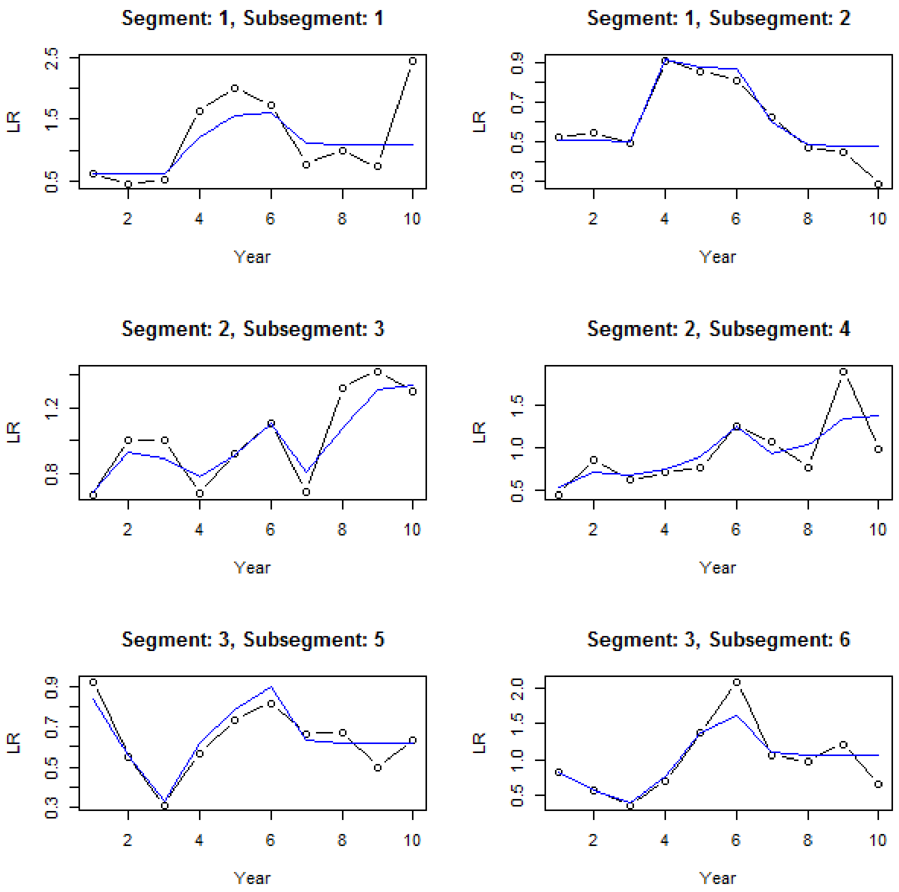
The code used to generate and fit the data is shown in Appendix B. The regression formula used was as follows, where a colon is used to indicate interaction effects, *ult.lr* are the ultimate loss ratios, *intercept* is an intercept term, *seg* is the variable for the segment, *subseg* is the variable for the sub-segment, and *yr.rw* is a random walk variable:

$$\text{log(ult.lr) = intercept + seg + subseg + yr.rw + seg:yr.rw + subseg:yr.rw}$$

This is a hierarchical credibility model. The overall average level of the loss ratio is determined by the intercept. The average relativities for each segment are determined by the *seg* coefficients, and the additional relativities for each sub-segment are determined by the *subseg* coefficients. Since each coefficient value is penalized and pushed back towards zero, each level is credibility weighted back towards the previous, i.e., each segment is credibility weighted back towards the overall mean and each sub-segment is credibility weighted back towards the segment. The same can be said for the changes by year. The *yr.rw* variable creates a random walk on the intercept, which affects all segments and sub-segments. The interaction of this variable and the segment allows for additional changes that only affect particular segments, and these changes are credibility weighted back towards the overall changes by year. The same occurs at the sub-segment level, and these changes are credibility weighted back towards the indicated segment changes.

The model just discussed does a good job of fitting all segments and their changes by year but does not take into account any possible autocorrelation between years, i.e., momentum of the yearly changes. Stated another way, if a change is observed in a previous period, then instead of using no change as the complement of credibility for the next period, perhaps the complement should be set to somewhere in between zero and the previous indicated change. This can be tested by redefining the random walk variable with different amounts of momentum and then refitting the model. The momentum value having the lowest cross validated error is then selected. The cross validated error (using the deviance as the error measure) for the model with no momentum is 0.0999. Testing in increments of 0.1, the next value tested is 0.1. This model has a cross validated error of 0.0984, which is better than the first model. Testing a momentum value of 0.2 yields an error of 0.1015, worse than the previous. So the selected value is 0.1. The final fitted results for each segment and sub-segment are shown in Figure 11.

Figure 11: Fitted trended, ultimate loss ratios by segment



Looking at this figure, some common trends can be seen. Most of the sub-segments start increasing at year 3 and decrease in year 7, although this increase is delayed a year in sub-segment 3. Common patterns can be seen by segment as well. Sub-segments in segment 1 show an initial increase followed by a decrease, segment 2 shows a generally increasing pattern, and segment 3 shows a decrease followed by an increase, and then no change from year 7. (All of this can be seen by looking directly at the fitted coefficients as well.) It is also apparent that less weight is given to the more recent years due to the additional uncertainty of these immature years.

10. CONCLUSION

This paper showed a method for incorporating a subset of state space model functionality into a linear regression framework. Besides for the improved performance and ease of use of this method, the resulting models are intuitive and the corresponding parameters lend themselves easily to interpretation. This can be a useful tool when attempting to "dig deeper" and discover changes or trends that may be affecting particular segments or entities. It can make forecasts into future periods more accurate as well. The results are suitable for presentation, which is an important consideration since findings often need to be communicated to other parties. Lastly, it is incorporated in a framework that scales well to large datasets, an important consideration in the age of big data.

APPENDIX A: Simulation Code

```
library( glmnet )
library( mgcv )
library( compiler )
library( rstan )

# logit functions for later
logit <- function(x) log( x / ( 1 - x ) )
ilogit <- function(x) exp(x) / ( 1 + exp(x) )

# function for creating dummy variables that include a column for every single value (unlike the default behavior in GLMs that leave out one value).
# this is needed for penalized regression (i.e. credibility models), since every segment is credibility weighted back towards the intercept (mean)
contr.pen <- function( n, contrasts=TRUE, sparse=FALSE ) contr.treatment( n, contrasts=contrasts, sparse=sparse )
contr.pen.sparse <- function( n, contrasts=TRUE, sparse=TRUE ) contr.treatment( n, contrasts=contrasts, sparse=sparse )

# function for creating dummy variables for implementing random walks
contr.randwalk <- function( n, contrasts = TRUE, sparse = TRUE, momentum=0, rel.cred=1, stdize=TRUE ) {
  if (length(n) <= 1L) {
   if (is.numeric(n) && length(n) == 1L && n > 1L)
      levels <- seq_len(n)
   else stop("not enough degrees of freedom to define contrasts")
  } else { levels <- n }
  levels <- as.character(levels)
  if ( sparse ) {
    cont <- Matrix( c(0), nrow=length(levels), ncol=length(levels) - 1, sparse=TRUE )
  } else {
    cont <- matrix( c(0), nrow=length(levels), ncol=length(levels) - 1 )
  }
  for ( i in 2:n ) {
    cont[, i - 1] <- ifelse( 1:n < i, 0, ifelse( rep( momentum, n ) == 1, ( 1:n - i + 1 ), ( 1 - momentum ^ ( 1:n - i + 1 ) ) / ( 1 - momentum ) ) )
    cont[, i - 1] <- cont[, i - 1] - mean( cont[, i - 1] )
  }
  if (contrasts) {
   colnames(cont) <- levels[-1]
  }
  # standardize
  if ( stdize ) {
    for ( i in 1:ncol(cont) ) {
      cont[,i] <- cont[,i] / sum( diff( cont[,i] ) ^ 2 ) ^ 0.5
    }
  }

  cont <- cont * rel.cred
  cont
}

# fit kalman filter
kf.fit <- function( x.obs, prem=rep( 1, length(x.obs) ), incl=1:length(x.obs) ) {
  kf <- cmpfun( function( params, return.values=F ) {
    n <- length(x.obs)
```

```
        x.pred0 <- rep( 0, n )
        x.pred <- rep( 0, n )
        k <- rep( 0, n )
        p0 <- rep( 0, n )
        p <- rep( 0, n )
        f <- rep( 0, n )
        err <- rep( 0, n )

        Q.est <- exp( params[1] ) * ilogit( params[2] )
        R.est <- exp( params[1] ) * prem[1] * ( 1 - ilogit( params[2] ) )

        for ( i in 1:n ) {
          x.pred0[i] <- ifelse( i > 1, x.pred[i - 1], params[3] )
          err[i] <- x.obs[i] - x.pred0[i]
          p0[i] <- ifelse( i > 1, p[i - 1] + Q.est, 0 )
          f[i] <- p0[i] + R.est / prem[i]
          k[i] <- ifelse( ( ! i %in% incl ) | f[i] == 0, 0, p0[i] / f[i] )
          p[i] <- p0[i] * ( 1 - k[i] )
          x.pred[i] <- x.pred0[i] + k[i] * err[i]
        }
        loglik <- sum( sapply( 1:n, function(i) log( max( 1e-100, dnorm( err[i], 0, f[i] ^ 0.5 ) ) ) ) )
        if ( return.values ) {
          x.smooth <- rep( 0, n )
          x.smooth[n] <- x.pred[n]
          for ( i in (n - 1):1 ) x.smooth[i] <- x.pred[i] + ( p[i] / p0[i + 1] ) * ( x.smooth[i + 1] - x.pred[i] )
          list( loglik=loglik, Q=Q.est, R=R.est, x=x.smooth, x.pred=x.pred, f=f, k=k, p=p )
        } else {
          loglik
        }
      } )
      o <- optim( c( log( var(x.obs) / 2 ), -10, sum( prem * x.obs ) / sum( prem ) ), kf, control=list( fnscale=-1 ),
method='BFGS' )
      kf( o$par, return.values=T )
    }

    # calculate root mean square error
    rmse <- function( x, y, i=1:length(x) ) mean( ( x[i] - y[i] ) ^ 2 ) ^ 0.5

    # number of years
    n <- 10
    # innovation variance
    Q <- 0.002
    # volaility
    R <- 0.003
    # serial correlation
    years.corr <- 0.1

    # definitions for no outliers
    q.large.pct.change <- 0
    r.large.pct.change <- 0
    q.df.reg <- 250
    q.df.large <- 250
    r.df.reg <- 250
    r.df.large <- 250
    q.large.pct.change <- 0
    r.large.pct.change <- 0
    pct.change <- 0.5
```

```
Q.large <- 0
R.large <- 0

# definitions for with outliers
q.df.reg <- 6
q.df.large <- 3
r.df.reg <- 12
r.df.large <- 6
q.large.pct.change <- 0.05
r.large.pct.change <- 0.1
pct.change <- 0.5
Q.large <- Q * 5
R.large <- R * 5

num.iter <- 250

do.graph1 <- FALSE
print.cvm1 <- FALSE
do.bayes <- TRUE
# run the simulation
results <- sapply( 1:num.iter, function(iter) {
  if ( iter %% 10 == 0 ) print( iter )

  # simulate data
  q <- rt( n, q.df.reg ) * Q ^ 0.5
  # add serial correlation
  for ( i in 2:n ) {
    q[i] <- years.corr * q[i - 1] + sqrt( 1 - years.corr ^ 2 ) * q[i]
  }
  q <- ifelse( runif( n ) < pct.change, q, rep( 0, n ) )
  q.large <- ifelse( runif( n ) < q.large.pct.change, rt( n, q.df.large ) * Q.large ^ 0.5, rep( 0, n ) )
  q <- q + q.large
  q.cuml <- cumsum( q )
  a <- 20 * exp(q.cuml)
  r.large <- ifelse( runif( n ) < r.large.pct.change, rt( n, r.df.large ) * R.large ^ 0.5, rep( 0, n ) )
  o <- exp( log( a ) + rt( n, r.df.reg ) * R ^ 0.5 + r.large )

  # create the time series variables
  d <- data.frame( y=1:n, a=a, o=o )
  # standardize trend variable
  d$y.std <- d$y / sqrt( n - 1 )
  d$y.std <- d$y.std - mean( d$y.std )
  d$y.fac <- factor( d$y )
  contrasts( d$y.fac ) <- contr.pen( nrow(d), sparse=TRUE )
  d$y.rw <- d$y.fac
  contrasts( d$y.rw ) <- contr.randwalk( nrow(d), sparse=TRUE )
  d$y.rw.mom <- d$y.fac
  contrasts( d$y.rw.mom ) <- contr.randwalk( nrow(d), momentum=0.25, sparse=TRUE )

  if ( do.graph1 ) {
    plot( d$y, d$o, type='b', xlab='Year', ylab='X', lwd=2 )
    lines( d$y, d$a, type='b', lty=3 )
    abline( h=mean( d$o ), col='gray' )
  }

  # function to fit models
  fit.model <- function( form, num.folds=3, num.times=20, alpha=0.75, fit.incl=1:n, do.graph=do.graph1,
print.cvm=print.cvm1, col='blue', lty=1, lwd=1 ) {
          x <- sparse.model.matrix( form, d )
         # remove the intercept (first column) since the penalized regression function adds it automatically
          # (but create the matrix with the intercept, since sometimes the model.matrix function won't use the correct
contrasts for the first term of the formula without an intercept)
          x <- x[, -1]

          fit <- glmnet( x[fit.incl,], log( d$o[fit.incl] ), family='gaussian', standardize=FALSE, alpha=alpha )    #,
lambda.min.ratio=1e-10 )
          fit.cv <- lapply( 1:num.times, function(i) cv.glmnet( x[fit.incl,], log( d$o[fit.incl] ), family='gaussian',
standardize=FALSE, alpha=alpha, nfolds=num.folds, lambda=fit$lambda ) )
          lambda <- mean( sapply( fit.cv, function(x) x$lambda.min ) )
          cvm <- mean( sapply( fit.cv, function(x) min( x$cvm ) ) )
          p <- exp( predict( fit, newx=x, type='response', s=lambda ) )
          if ( do.graph ) lines( d$y, p, col=col, lwd=lwd, lty=lty )
          if ( print.cvm ) print( cvm )
          p
        }

        d$p.mean <- mean( d$o )
        d$p.fac <- fit.model( ~ y.fac, col='blue', print.cvm=print.cvm1 )
        d$p.rw <- fit.model( ~ y.rw, col='red', lwd=2, print.cvm=print.cvm1 )
        d$p.rw.mom <- fit.model( ~ y.rw.mom, col='red', lty=2, lwd=2, print.cvm=print.cvm1 )

        # kalman filter
        d$p.kf <- exp( kf.fit( log( d$o ), incl=1:n )$x )
        if ( do.graph1 ) lines( d$y, d$p.kf, col='green' )

        # kalman filter with bagging
        fit.kf.bag <- rowSums( exp( sapply( 1:10, function(i) kf.fit( log( d$o ), incl=sample( 1:n, round( n * (2/3) ) ) )$x ) ) )
/ 10
        d$p.kf.bag <- fit.kf.bag
        if ( do.graph1 ) lines( d$y, d$p.kf.bag, col='green', lty=2 )

        # spline
        fit.sp <- gam( log( o ) ~ s( y ), data=d )
        d$p.sp <- exp( predict( fit.sp, newdata=d ) )
        if ( do.graph1 ) lines( d$y, d$p.sp, col='gold', lwd=2 )

        # bayesian model
        if ( do.bayes ) {
          # model on log of observed instead of handling in model, so that comparable to other models
          data.stan <- list(
            x=log( d$o )
            , N=n
            , y_rw_scale=5
          )
          fit.bayes <- stan( 'rw_test.stan', data=data.stan, chains=3, iter=1000 )
          stan.obj <- extract( fit.bayes, permuted=TRUE )
          #traceplot( fit.bayes )
          #stan_dens( fit.bayes, separate_chains=TRUE )
          d$p.bayes <- exp( sapply( 1:n, function(i) mean( stan.obj[['y']][,i] ) ) )
          if ( do.graph1 ) lines( d$y, d$p.bayes, col='pink', lwd=2 )
        }

        # calculate prediction errors of each method
        rmse.pred <- c()
```

```
  for ( f in names(d)[ grep( 'p.', names(d), fixed=TRUE ) ] ) {
    rmse.pred <- c( rmse.pred, rmse( d[[f]], d$a, 1:n ) )
    names( rmse.pred )[length(rmse.pred)] <- f
  }
  rmse.pred
} )

data.frame( method=rownames( results ), rmse=sapply( 1:nrow(results), function(i) mean( results[i,] ) ) )
```

Save in separate file names "rw_test.stan":

```
data {
  int N;
  vector<lower=0>[N] x;
  real<lower=0> y_rw_scale;
}

parameters {
  vector<lower=-2, upper=2>[N - 1] y_rw;
  real<lower=0, upper=2> y_sd;
  real<lower=0> y_rw_sd;
  real y1;
}

transformed parameters {
  vector[N] y;

  y[1] = y1;
  for ( i in 2:N ) {
    y[i] = y[i - 1] + y_rw[i - 1];
  }
}


model {
  y1 ~ uniform( 1, 5 );
  y_rw_sd ~ cauchy( 0, y_rw_scale );
  y_rw ~ normal( 0, y_rw_sd );
  x ~ normal( y, y_sd );
}
```

APPENDIX B: Loss Ratio Example

```
library( HDtweedie ) # elastic net for Tweedie family
library( Matrix ) # for building sparse matrices
library( ggplot2 )

# function for creating dummy variables that include a column for every single value (unlike the default behavior in
GLMs that leave out one value).
# this is needed for penalized regression (i.e. credibility models), since every segment is credibility weighted back
towards the intercept (mean)
contr.pen <- function( n, contrasts=TRUE, sparse=FALSE ) contr.treatment( n, contrasts=contrasts, sparse=sparse
)
contr.pen.sparse <- function( n, contrasts=TRUE, sparse=TRUE ) contr.treatment( n, contrasts=contrasts,
sparse=sparse )

# function for creating dummy variables for implementing random walks
contr.randwalk <- function( n, contrasts = TRUE, sparse = TRUE, momentum=0, rel.cred=1, stdize=TRUE ) {
  if (length(n) <= 1L) {
    if (is.numeric(n) && length(n) == 1L && n > 1L)
        levels <- seq_len(n)
    else stop("not enough degrees of freedom to define contrasts")
  } else { levels <- n }
  levels <- as.character(levels)
  if ( sparse ) {
     cont <- Matrix( c(0), nrow=length(levels), ncol=length(levels) - 1, sparse=TRUE )
  } else {
     cont <- matrix( c(0), nrow=length(levels), ncol=length(levels) - 1 )
  }
  for ( i in 2:n ) {
     cont[, i - 1] <- ifelse( 1:n < i, 0, ifelse( rep( momentum, n ) == 1, ( 1:n - i + 1 ), ( 1 - momentum ^ ( 1:n - i + 1 ) )
/ ( 1 - momentum ) ) )
     cont[, i - 1] <- cont[, i - 1] - mean( cont[, i - 1] )
  }
  if (contrasts) {
    colnames(cont) <- levels[-1]
  }
  # standardize
  if ( stdize ) {
     for ( i in 1:ncol(cont) ) {
        cont[,i] <- cont[,i] / sum( diff( cont[,i] ) ^ 2 ) ^ 0.5
     }
  }

  cont <- cont * rel.cred
  cont
}

num.segs <- 3
num.subsegs.each <- 2
num.subsegs <- num.segs * num.subsegs.each
# for finding the appropriate segment for each subsegment
seg.map <- c( sapply( 1:num.segs, function(i) rep( i, num.subsegs.each ) ) )
num.yrs <- 10
years.corr <- 0.2
ldf <- 3 ^ ( 0.65 ^ ( ( num.yrs - 1 ):0 ) )
```

```
  # --- simulate losses ---
  overall.avg <- log( 700 )
  # segment relativities
  seg.rel <- rnorm( num.segs, 0, 0.1 )
  # subsegment relativies
  subseg.rel <- rnorm( num.subsegs, 0, 0.1 )
  # random walk coefs
  rw.chg.init <- rnorm( num.yrs, 0, 0.2 )
  rw.seg.chg.init <- sapply( 1:num.segs, function(i) rnorm( num.yrs, 0, 0.15 ) )
  rw.subseg.chg.init <- sapply( 1:num.subsegs, function(i) rnorm( num.yrs, 0, 0.1 ) )
  # add correlation (momentum) to the changes
  rw.chg <- rw.chg.init
  rw.seg.chg <- rw.seg.chg.init
  rw.subseg.chg <- rw.subseg.chg.init
  for ( i in 2:num.yrs ) {
    rw.chg[i] <- years.corr * rw.chg.init[i - 1] + sqrt( 1 - years.corr ^ 2 ) * rw.chg.init[i]
    for ( j in 1:num.segs ) rw.seg.chg[i, j] <- years.corr * rw.seg.chg.init[i - 1, j] + sqrt( 1 - years.corr ^ 2 ) * rw.seg.chg.init[i,
j]
    for ( j in 1:num.subsegs ) rw.subseg.chg[i, j] <- years.corr * rw.subseg.chg.init[i - 1, j] + sqrt( 1 - years.corr ^ 2 ) *
rw.subseg.chg.init[i, j]
  }

  # using simulated values, produce the data, with error. (note this is a crude simulation)
  for ( subseg in 1:num.subsegs ) {
    seg <- seg.map[subseg]
    x <- exp( overall.avg + seg.rel[seg] + subseg.rel[subseg]
      + cumsum( rw.chg ) + cumsum( rw.seg.chg[,seg] ) + cumsum( rw.subseg.chg[,subseg] )
      + rnorm( num.yrs, 0, 0.2 * ldf ) ) / ldf
    d.add <- data.frame( seg=seg, subseg=subseg, yr=1:num.yrs, loss=x, ldf=ldf, ep=1000 )
    if ( subseg == 1 ) {
      d <- d.add
    } else {
      d <- rbind( d, d.add )
    }
  }
  d$seg <- factor( d$seg )
  d$subseg <- factor( d$subseg )

  # cape cod-like method to develop losses and weight years
  d$ult.lr <- d$loss * d$ldf / d$ep
  # this will give the greener years less weight
  d$used.ep <- d$ep / d$ldf

  # --- build the model ---
  # create random walk variable
  # set different values of the momentum (and relative credibility if want) here
  rw.momentum <- 0
  rw.rel.cred <- 1
  # rw.momentum <- 0.1
  d$yr.rw <- factor( d$yr )
  contrasts( d$yr.rw ) <- contr.randwalk( num.yrs, sparse=TRUE, momentum=rw.momentum, rel.cred=rw.rel.cred )

  # create the modeling  matrix that describes the independent variables of the model (use a sparse matrix)
  # change the default contrasts (i.e. dummy variable encodings) to create columns for every level of a variable
  options( contrasts = c( 'contr.pen.sparse', 'contr.pen.sparse' ) )
  # (no intercept in formula since glmnet function adds automatically)
  x <- sparse.model.matrix( ~ seg + subseg + yr.rw + seg:yr.rw + subseg:yr.rw, d )
```

```
# remove the intercept (first column) since the penalized regression function adds it automatically
# (but create the matrix with the intercept, since sometimes the model.matrix function won't use the correct contrasts for the first term of the formula without an intercept)
x <- x[, -1]
options( contrasts = c( 'contr.treatment', 'contr.poly' ) )

# fit the model
# (HDtweedie doesn't support sparse matrices unlist glmnet, so need to convert back to unsparse matrix)
fit <- HDtweedie( as.matrix( x ), d$ult.lr, weights=d$used.ep, p=1.9, alpha=0.75, standardize=FALSE, lambda.factor=1e-3 )
# cross validate to get the optimal penalty parameter
set.seed( 1112 )
fit.xval <- cv.HDtweedie( as.matrix( x ), d$ult.lr, weights=d$used.ep, p=1.9, alpha=0.75, standardize=FALSE, lambda.factor=1e-3, nfolds=5 )
# cross validated deviance (error) of the model - use this to test the momentum parameter
min( fit.xval$cvm )
lambda <- fit.xval$lambda.min
# check that the chosen penalty isn't at the ends
which( lambda == fit.xval$lambda )
length( fit.xval$lambda )

# make predictions
d$ult.lr.pred <- predict( fit, as.matrix( x ), type='response', s=lambda )[,1]

# --- graph results ---
# one at a time
par( mfrow=c( 3, 2 ) )
for ( subseg in 1:num.subsegs ) {
  plot( 1:num.yrs, d$ult.lr[d$seg == seg.map[subseg] & d$subseg == subseg], type='b', xlab='Year', ylab='LR'
    , main=paste( 'Segment: ', seg.map[subseg], ', Subsegment: ', subseg, sep='' ) )
  lines( 1:num.yrs, d$ult.lr.pred[d$seg == seg.map[subseg] & d$subseg == subseg], col='blue' )
}
par( mfrow=c( 1, 1 ) )


d.plot <- rbind( data.frame( Year=d$yr, seg=d$seg, subseg=d$subseg, Segment=paste( d$seg, d$subseg ), LR=d$ult.lr, Line='Actual' )
    , data.frame( Year=d$yr, seg=d$seg, subseg=d$subseg, Segment=paste( d$seg, d$subseg ), LR=d$ult.lr.pred, Line='Predicted' ) )

# each segment together
seg <- 1
ggplot( d.plot[d.plot$seg == seg,], aes( Year, LR ) ) + geom_line( aes( color=Segment, linetype=Line ) )

# all together
ggplot( d.plot, aes( Year, LR ) ) + geom_line( aes( color=Segment, linetype=Line ) )
```

10. REFERENCES

[1] Carlin, B. 1992. State Space Modeling of Non-Standard Actuarial Time Series. Insurance: Mathematics and Economics. October 1992. Volume 11, Issue 3. pp 209-222.

[2] De Jong, P. 2005. State Space Models in Actuarial Science. Macquarie University Actuarial Studies, Research Paper. No. 2005/02. July 2005.

[3] De Jong, P. and Zehnwirth, B. 1983. Claims Reserving, State-Space Models and the Kalman Filter. Journal of the Institute of Actuaries, 110, pp 157-181.

[4] Dempster, A. P., Laird N. M., and Rubin D. B. 1977. Journal of the Royal Statistical Society. Series B (Methodological). Vol. 39, No. 1 (1977), pp. 1-38
[5] Evans, Jonathan P. and Frank Schmid. 2007. Forecasting Workers Compensation Severities and Frequency Using the Kalman Filter. Casualty Actuarial Society Forum, 2007: Winter, pp 43–66
[6] Frees, E. and Gee, L. 2016. Rating Endorsements Using Generalized Linear Models. Variance 10:1, 2016, pp 51-74
[7] Friedman, J., Hastie, T., and Tibshirani, R. 2009. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1):1, 2010
[8] Gelman, A. 2008. Scaling regression inputs by dividing by two standard deviations. Stat. Med. 27: 2865–2873.
[9] Geweke, J. 1977. The Dynamic Factor Analysis of Economic Time Series. Latent Variables in Socio-Economic Models. Amsterdam, North Holland.
[10] Hastie, T., Tibshirani, R., and Friedman, J. 2009. Elements of Statistical Learning, Volume 2, New York: Springer, 2009
[11] Hastie, T. and Qian, J. 2014. Glmnet Vignette. http://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html
[12] Kim C. and Nelson C. 1999. State-Space Models with Regime Switching. MIT Press, 1999. pp. 29-30, 36-37
[13] Korn, U. 2016. An Extension to the Cape Cod Method with Credibility Weighting Smoothing. Casualty Actuarial Society E-Forum, 2016: Summer, pp 5–27
[14] Taylor, G. and McGuire, G. 2007. Adaptive Reserving using Bayesian revision for the Exponential Dispersion Family. Centre for Actuarial Studies, Department of Economics, University of Melbourne.
[15] Williams, B., Hansen, G., Baraban, A., and Santoni, A. 2015. A Preactical Approach to Variable Selection – A Comparison of Various Techniques. Casualty Actuarial Society E-Forum, 2015: Summer, pp 4-40
[16] Wüthrich, M.V. and Merz, M. 2008. Stochastic Claims Reserving Methods in Insurance. Wiley.
[17] Zehnwirth B. 1996. Kalman Filters with Applications to Loss Reserving. Research Paper Number 35, Centre for Actuarial Studies, The University of Melbourne.
[18] Zou, H. and Hastie, T. 2005. Regularization and Variable Selection via the Elastic Net. J. Royal. Stat. Soc. B 2005;67(2), pp 301-320

Biography of the Author

Uri Korn is a Director of Predictive Modeling for AIG, Client Risk Solutions where he uses advanced analytics to assist clients in reducing their losses. Prior to that, he was an AVP & Actuary at Axis Capital serving as the Research and Development support for all commercial lines of insurance. He has published papers on non-aggregated loss development techniques, time series, and several papers on practical approaches to credibility. Recently, he was awarded the 2017 Ratemaking Prize for the best call paper. He graduated from the University of Pennsylvania in 2003 with a BSE in Computer Science in Engineering and is a Fellow of the Casualty Actuarial Society.

The Average Maturity of Loss Approximation of Loss Development

By

Ira Robbin, PhD

Abstract

This paper will present a formula for generalizing the average date of loss approximation (ADOL) so it operates reasonably at immature ages, ages where the usual ADOL approximation breaks down. The formula adjusts the evaluation date on the approximating exposures so they have the same average maturity of loss (AMOL) as those being approximated. The formulas also accounts for differences in the percent of exposure to loss that has been accrued at the respective ages. The proof of the formula is shown. It is based on decomposing total loss development for a set of exposures into separate pure loss development and exposure bucketing terms. The derivation also leads to an error term that can be used to gauge the accuracy of the approximation. The paper features examples in which the AMOL is applied to approximate policy year and policy year cut-off loss development.

Keywords Loss Development Factors, Average Date of Loss, Average Maturity of Loss

1. INTRODUCTION

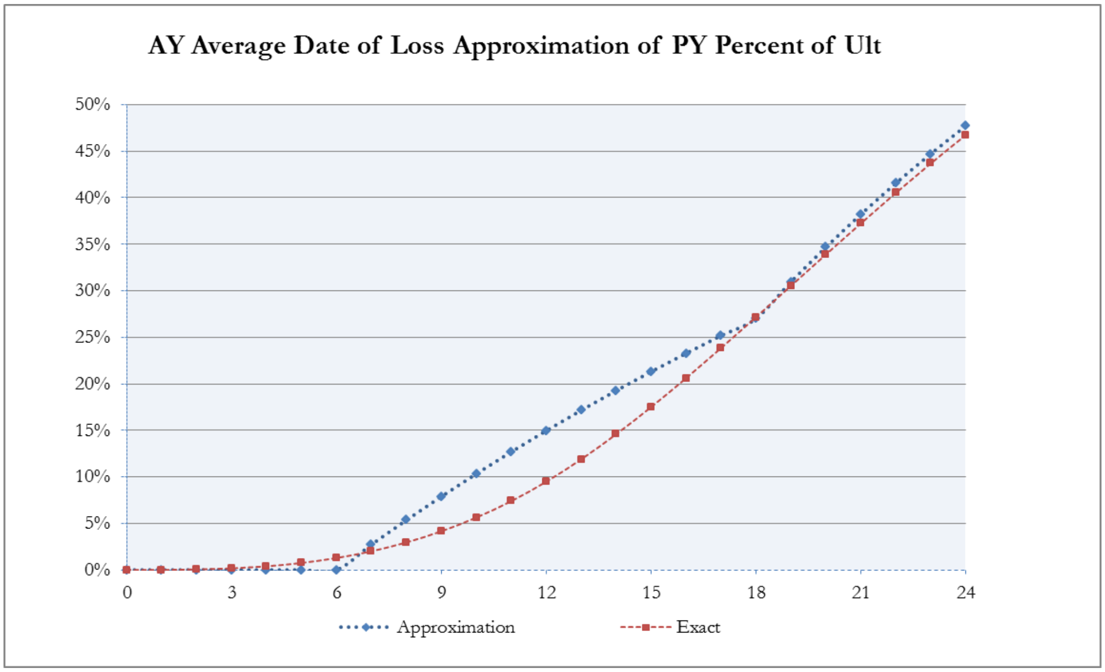
Many property and casualty actuaries have at some point been asked to estimate policy year (PY) loss development patterns when the only available loss development factors (LDF) were on an accident year (AY) basis. Rather than exit the spreadsheet saying that the differences between the policy year and accident year exposure patterns make the task impossible, actuaries have instead looked at the problem more carefully and arrived at the conclusion that the available information should be sufficient to obtain a fair approximation. Intuitively, they realize the AY development pattern encodes enough information so that the actuary should be able to take the AY pattern, make appropriate adjustments, and arrive at a good estimate of PY loss development. The question then is: what exactly are these appropriate adjustments?

The standard answer to this question is that the evaluation date must be adjusted for the difference in the average date of loss. This produces the Average Date of Loss (ADOL) approximation. Under the ADOL approximation, the policy year LDF at a given age is estimated by the accident year LDF six months earlier. The six month adjustment reflects the difference between the six month average date of loss for a uniform accident year and the twelve month average date of loss for a policy year.

Though the ADOL method works reasonably well for mature ages, it breaks down when exposures are immature. In particular the usual ADOL approximation of a policy year starts to go awry for ages less than 24 months and makes no sense for ages of 6 months or less.

The graph in Figure 1 is an example in which the ADOL approximation is graphed against the true policy year loss development percent of ultimate curve. The numbers behind the graph are derived and presented in Exhibit 2. The error is significant for ages below 15 months. The approximation at age 12 is 14.96%, but the actual percentage is 9.48% and the estimated Age-to-Ultimate is 6.685 but the actual is 10.547. Of course, one should not even be using this approximation at an age where it is invalid, but some have done so for want of an alternative without having a good sense of how significant an error might be introduced.

Figure 1



How can one obtain an approximation that works at earlier ages, when the exposures are not fully earned? The purpose of this paper is to provide an answer to that question. This paper will present an approximation, the Average Maturity of Loss (AMOL) which generalizes the ADOL. The formula extends to immature ages and the derivation also produces an equation for the error term. The AMOL method provides a general approach for estimating the LDF for one set of exposures based on the LDF from another. It reflects two adjustments:

- **Age Adjustment**: The evaluation age for the LDF of the known curve is adjusted so both LDF are for losses with same conditional average maturity of loss. For example, a policy year at 12 months has conditional average date of loss equal to 8 months and thus a 4 month conditional average maturity of loss. An accident year evaluated at 8 months also has a 4 month conditional average maturity.
- **Earned Exposures Adjustment:** A ratio needs to be applied to account for differences in the cumulative earned exposures. For example, on a policy year as of 12 months the cumulative exposures are at 50%, while those for an accident year as of 8 months are at 67%. Thus the PY LDF as of 12 months can be approximated by the AY LDF as of 8 months times the ratio, 67/50.

The general AMOL uses one set of LDF to estimate those for another set of exposures. It is the logical consequence of an even more basic idea that development for any set of exposures is approximable as the product of two terms. One term reflects is the percent of exposures earned as of the evaluation date. The other is the percent of development on losses for a length of time equal to the conditional average maturity of loss for that set of exposures. The general AMOL formula will be proved using the Taylor series expansion on the loss development integral representation previously presented in Robbin [3] and Robbin and Homer [4]. An expression for the error will be included in the result. Note the use of conditional average maturities and the cumulative percentage adjustment allow the formula to apply at immature ages. A conceptually similar two-factor approach was presented by David Clark in Appendix B of his paper [2] on LDF curve fitting. The formula will therefore sometimes be referred to as the *Clark-Robbin Two-Factor AMOL Formula.*

The paper will provide a numerical example in which AY LDF are used to approximate LDF for a policy year and a cut-off policy year using the standard ADOL and the generalized AMOL.. The need to develop a cut-off policy year arises in evaluating treaty year experience in reinsurance. Some risks attaching contracts allow the cedant to cut off the unearned exposures at the end of the first year. What is left is equivalent to half a policy year. In this example the exact development patterns will be constructed from the Robbin and Homer [4] generating formula. This allows computation of the exact factors for both curves and thus an explicit calculation of the error of the approximations.

In Chapter 2 the Robbin and Homer perspective on loss development will be presented. This cleanly separates the inherent lag in how long it takes for claims to be reported and to be settled from the way that claims are bucketed for accounting and evaluation purposes. Chapter 3 will contain the statement and proof of the general AMOL Approximation and error term. Chapter 4 will show the AMOL approximation on policy year and policy year cut-off exposures.

In summary, the general AMOL method presented in this paper will allow actuaries to extend the ADOL formula to early ages and it will also provide a solid practical framework for approximating development for less common exposure groupings.

2. LOSS DEVELOPMENT FROM CLAIMS TO TRIANGLES

Under the perspective found in Robbin and Homer [4] and Robbin [3] there are inherent stochastic processes operating at the individual claim level that describe individual claim development. These quantify how long it takes for the claim to be reported after it occurs, how its valuation changes over time, and how long it takes to settle. In going from individual claims to aggregated triangles, a bucketing is done that assigns loss activity on designated sets of claims over designated evaluation dates to specific cells. Under this perspective, development of a loss triangle is the result of both exposure bucketing and underlying claims development. Though this seems very theoretical, it leads to a useful general mathematical representation of loss development.

Following Robbin and Homer [4] and Robbin [3] but slightly revising the notation, let T be the underlying claim development lag random variable defined as the time elapsed from when a claim occurs until when a unit amount of development activity gets recorded. The development activity could be the recording of a reported claim, the settlement of a claim, the payment of a claim, or any other variable of interest usually arrayed in triangular format.

Let A be a loss exposure bucketing random variable defined as the lag from the start of an exposure period until a loss occurs. For an accident year under the usual assumptions, A is uniform on [0, 1]. Let $PCT_{T|A}$ (t) be the percent of ultimate at age t for the underlying development variable, T, and the exposure and evaluation bucketing variable, A.

Assuming all losses are subject to the same underlying loss development process, the percent of ultimate function can be expressed as the convolution integral:

Robbin-Homer Convolution Formula for Percent of Ultimate (2.1)

$$PCT_{T|A}(t) = F_{A+T}(t) = \int_0^t ds\, f_A(s) * F_T(t-s)$$

The integral representation assumes the random variables A and T are independent. Independence can be asserted based on the general grounds that the manner in which loss exposures are bucketed for purposes of accounting and reporting should not have any impact on how the claims are settled.

Using Equation 2.1, one can derive an approximation in which separation between development and grouping is expressed mathematically with terms and factors that depend either on the underlying development or on the grouping of exposures. It is first useful to define several mathematical terms:

Definitions (2.2)

2.2.1. $Conditional\ Average\ Date\ of\ Loss\ Exposure\colon\ m_A(t) = E_A[T|T < t]$

2.2.2. $Conditional\ Average\ Maturity\ of\ Loss\ Exposure\colon\ r_A(t) = t - m_A(t)$

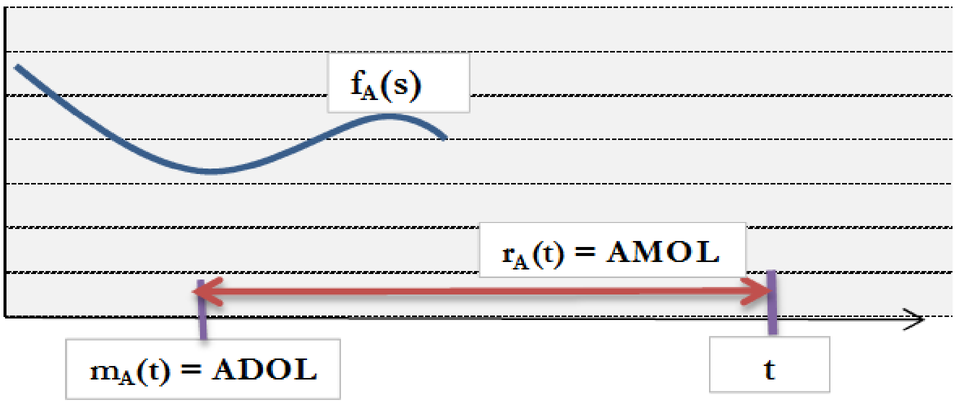
2.2.3. $Conditional\ Variance\ of\ Date\ of\ Loss\ Exposure\colon\ v_A(t)$
$= E_A[(T - m_A(t))^2|T < t]$

Figure 2 shows the relation between the average date of loss and the average maturity of loss.

Formulas for the average date of loss, average maturity of loss, and average variance of loss date are shown in the Appendix for an accident year, a Cut-off Policy Year, and a Policy year.

Figure 2

Average Date of Loss and Average Maturity of Loss



One can now approximate development using a two-factor formula. One first factor is the percentage of ultimate exposure at the evaluation date. The second is the percent of ultimate for the underlying loss development process at the average maturity of loss. This is the Clark-Robbin Two-Factor Average Maturity of Loss Approximation. The error in the approximation is given by a term that includes the variance in the loss exposure date. The percent of ultimate of losses subject to development T when bucketed as per exposure distribution A is approximated via follows:

Clark-Robbin Two-Factor Average Maturity of Loss Approximation (2.3)

$$PCT_{T|A}(t) \approx F_A(t) \cdot PCT_T\big(r_A(t)\big)$$

$$with\ error\ F_A(t) \cdot \left\{\tfrac{1}{2} \cdot f_T'\big(r_A(\tau)\big) \cdot v_A(t)\right\}\ where\ 0 < \tau < t$$

Proof:

Apply the Taylor series expansion up to second order to write:

$$F_T(t-s) = F_T\big(t - m_A(t)\big) \\ - \big(s - m_A(t)\big)f_T\big(r_A(t)\big) + \tfrac{1}{2}\big(s - m_A(t)\big)^2 f_T'\big(r_A(\tau)\big) \tag{2.4}$$

$$with\ 0 \le \tau \le t$$

Plugging this into the integral in 2.1 leads to:

$$PCT_{T|A}(t) = \int_0^t ds\, f_A(s) \cdot F_T\big(r_A(t)\big) - \int_0^t ds\, f_A(s) \cdot \big(s - m_A(t)\big) \cdot f_T\big(r_A(t)\big) \tag{2.5}$$

$$+\tfrac{1}{2} \cdot f_T\big(r_A(\tau)\big) \int_0^t ds\, f_A(s) \cdot \big(s - m_A(t)\big)^2$$

The integral with the term s- $m_A(t)$ is zero. So the expression simplifies to:

$$PCT_{T|A}(t) = F_T\big(r_A(t)\big)F_A(t) + \tfrac{1}{2} f_T'\big(r_A(\tau)\big)F_A(t)v_A(t) \tag{2.6}$$

The approximation in Equation 2.3 is obtained by taking the first term. This comprised of a product of two factors. The remaining term is the error. For many bucketing distributions, this second term is relatively small. If finer accuracy is desired and a particular functional form is assumed that allows computation of the derivative of the development density, f′, one can approximate the true error term by using $r_A(t)$ in place of $r_A(\tau)$.

In words, Equation 2.3 means that the percent of ultimate for underlying development variable T and bucketing variable A is given as the product of the cumulative distribution of the bucketing random variable time t multiplied by the percent of ultimate for the development variable at the average maturity at time t plus a second order correction.

For an explicit application of Equation 2.3, assume underlying development is given as an exponential with mean, μ, so that:

Exponential CDF, Density, and Derivative of Density (2.7)

- $F_T(t) = 1 - \exp(-t/\mu)$
- $f_T(t) = (1/\mu) \exp(-t/\mu)$
- $f'_T(t) = -(1/\mu^2) \exp(-t/\mu)$

Formulas for the conditional average date of loss, the conditional average maturity of loss, and conditional variance for an accident year are shown in Appendix A. It follows that for t<1 (years), percent of ultimate development is approximated via:

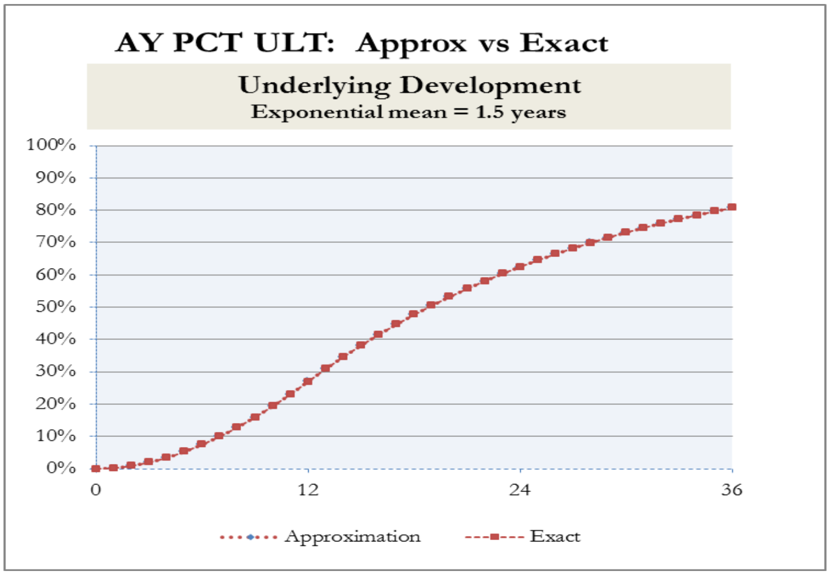
Bucketing-Development Approximation of Accident Year Development Based on Underlying Exponential Loss Development for t<1 (2.8)

$$PCT_{T|A}(t) \approx t \cdot \left\{ 1 - \exp\left(-\frac{t}{2\mu}\right) + \frac{-1}{24} \cdot \left(\frac{t}{\mu}\right)^2 \exp\left(-\frac{t}{2\mu}\right) \right\}$$

The formulas in 2.3, 2.7, and the Appendix can be used to derive a percent of ultimate approximate for t>1 for accident year development with an underlying exponential development process. This is left as an exercise for the reader. The graph in Figure 3 shows the approximation with the second order correction term by month out to month 36 for an exponential with mean equal to 1.50.

Exhibit 1 shows the derivation and values graphed in Figure 3. The approximation is extremely good as expected. The graph shows the curves are on top of one another. Exhibit 1 also shows the first order approximation and the reader can see that it, too, is quite good

Figure 3



. Another important and useful result is obtained when Equation 2.1 is applied to accident year exposures. In that case, Equation 2.1 can be expressed using formulas involving the limited expected value of T, denoted here as LEV:

Robbin Accident Year Percent of Ultimate Formula Based on LEVs (2.9)

$$PCT_{T|AY}(t) = \begin{Bmatrix} t - LEV(t) & for\ t < 1 \\ 1 - (LEV(t) - LEV(t-1) & for\ t > 1 \end{Bmatrix}$$

The proof is in Robbin [3]. Equation 2.9 provides a convenient way to generate accident year loss development curves given a parametric non-negative random variable such as a Pareto or exponential that has a tractable limited expected value formula. It was used to generate the exact AY curve used to approximate the policy year curve depicted in Figure 1.

3. THE GENERAL AVERAGE MATURITY OF LOSS APPROXIMATION

Now suppose A is an exposure bucketing random variable and assume a loss development curve based on A is known. Suppose B is another exposure bucketing random variable. The question arises: how can one approximate the development curve based on T given exposure B using the known percent of ultimate loss development curve based on T given exposure A? The answer is that a reasonable approximation using average maturity of loss can be obtained. In general, two adjustments need to be made:

- The evaluation date needs to be adjusted so the A-exposed losses at the adjusted date have the same average maturity as the B-grouped losses at the original evaluation date.
- There needs to be an adjustment for differences in the percent of exposures earned to date for A and B at the respective dates.

The General Average Maturity of Loss Approximation formula with error terms is given in Equation 3.1.

Generalized Average Maturity of Loss Approximation (3.1)

$$If\ r_B(t) = t - m_B(t) = r_A(t\,*) = \ t^* - m_A(t^*), then$$

$$PCT_{T|B}(t) = \frac{F_B(t)}{F_A(t^*)} \cdot PCT_{T|A}(t^*)$$

$$+\tfrac{1}{2} \cdot \left(v_B(t) - v_A(t^*)\right) \cdot f_T'(r_A(t^*))$$

Proof: The proof follows directly from Equation 2.3 and a little algebra and is left as an exercise for the reader.

The first step in applying 3.1 is to find the evaluation date t* that yields the same average maturity of loss:

Average Maturity Date Equation (3.2)

$$r_B(t) = \ r_A(t^*)$$

The next step requires an evaluation of the term, $PCT_{T|A}(t^*)$, the percent of ultimate for the known curve at the adjusted evaluation date. This will require either an explicit formula for the known curve at all evaluation dates or use of an interpolation routine.

4. ACCIDENT YEAR APPROXIMATION OF POLICY YEAR AND CUT-OFF POLICY YEAR LOSS DEVELOPMENT PATTERNS

In this section, the general AMOL approximation will be applied to estimate policy year and cut-off policy year development based on known development for an accident year. A cut-off policy year is one which includes claims on policies that are written uniformly over the year but which cuts off those that occur after the end of the year.[1]

The average maturity approximation requires finding for each evaluation date, t, the corresponding adjusted evaluation date, t*, so that the accident year has the same average maturity as the cut-off policy year.

Using the equations in the Appendix for the cut-off policy year and solving for t* as a function of t to satisfy 3.3, one finds:

AY Evaluation Date to Achieve Same Average Maturity as Cut-off Policy Year (4.1)

4.1.1 For $t \leq 1$, it follows that $t/3 = t^*/2$ which implies $t^* = (2/3)t$.

4.1.2. For $1 < t \leq 7/6$, it follows that $t - 2/3 = t^*/2$ which implies $t^* = 2t - 4/3$.

4.1.3. For $7/6 < t$, it follows that that $t - 2/3 = t^* - ½$ which implies $t^* = t - 1/6$.

This is demonstrated in Exhibit 3A. A key point is that the maturity age adjustments and loss exposure distribution adjustments by age do not depend at all on the loss development curves but only on the exposure bucketing curves. Note the adjustment varies over time. At 6 months, the cut-off policy year has a 4 month conditional average date of loss and a 2 month conditional average maturity of loss. The corresponding AY evaluation date is 4 months as that also leads to a 2 month average maturity of loss. When t is 12 months, the corresponding t* is 8 months as was previously observed. Later when t is 24 months, the average maturity of loss for the cut-off policy year is 16 months (24-8). The corresponding t* for an accident year at that stage is 22 months (22-6 =16).

Exhibit 3B shows how the general average maturity of loss approximation in Formula 3.1 is used to estimate a cut-off policy year using the accident year curve and where the

[1] Boor [1] discusses this briefly but describes it in terms of the first accident year for a company that just started writing at the start of the first year. The author's terminology borrows from a common clause in risks attaching reinsurance treaties that allows the cedant to cut-off the remaining unearned exposures at the end of the year.

underlying loss development lag is assumed to be an exponential distribution. The accident year curve was derived in Exhibit 1A. Exhibit 3B also shows the exact percent of ultimate and the age-to-ultimate LDF factor for the cut-off policy year along with the associated errors of the approximation. The reader is cautioned that the AMOL approximation is just that, an approximation. The reader is invited to verify that the errors are within the bounds specified by Equation 3.1.

The AMOL approximation of policy year LDF based on accident year patterns is shown in Exhibits 4A and 4B. The conditional average date of loss, average maturity of loss, and adjusted age are computed in Exhibit 4A. Policy year formulas are found in the Appendix. Exhibit 4B shows the evaluation at the AMOL adjusted evaluation date and the exposure ratio adjustment. Note after 24 months the AMOL approximation of policy year loss development become the same as the usual ADOL approximation.

5. CONCLUSION

Perhaps the immediate practical lesson is that use of the standard ADOL can seriously understate the Policy Year Age-to-Ultimate factors at ages 12 and 15. However, the AMOL approximation presented in this paper can be used to give a much more accurate, though certainly not exact, answer.

More generally the Average Maturity of Loss approximation has been shown to be a conceptually sound method for approximating loss development for an arbitrary bucketing of loss exposures over immature as well as mature periods. The Two-Factor formula makes good intuitive sense from a standard actuarial perspective and it is supported by a solid mathematical foundation.

Glossary of Exhibits

Exhibit 1A Robbin-Homer LEV Formula for AY Percent of Ultimate

AY PCT of Ultimate

$PCT_{T\|AY}(t)$ =	t-LEV(t)	t<1
	1-(LEV(t)-LEV(t-1))	t>1

		T = Exponential			
		mean	1.50	Exact Formula	
t months	t in years	LEV(t)	LEV(t-1)	AY PCT of ULT	AY ATU LDF
0	0.000				
1	0.083	0.081	0.000	0.23%	####
2	0.167	0.158	0.000	0.89%	####
3	0.250	0.230	0.000	1.97%	50.7033
4	0.333	0.299	0.000	3.44%	29.0365
5	0.417	0.364	0.000	5.29%	18.9163
6	0.500	0.425	0.000	7.48%	13.3695
7	0.583	0.483	0.000	10.00%	9.9952
8	0.667	0.538	0.000	12.84%	7.7859
9	0.750	0.590	0.000	15.98%	6.2580
10	0.833	0.639	0.000	19.40%	5.1556
11	0.917	0.686	0.000	23.08%	4.3330
12	1.000	0.730	0.000	27.01%	3.7020
13	1.083	0.771	0.081	30.96%	3.2303
14	1.167	0.811	0.158	34.69%	2.8828
15	1.250	0.848	0.230	38.22%	2.6166
16	1.333	0.883	0.299	41.56%	2.4064
17	1.417	0.917	0.364	44.71%	2.2364
18	1.500	0.948	0.425	47.70%	2.0963
19	1.583	0.978	0.483	50.53%	1.9791
20	1.667	1.006	0.538	53.20%	1.8796
21	1.750	1.033	0.590	55.73%	1.7943
22	1.833	1.058	0.639	58.12%	1.7205
23	1.917	1.082	0.686	60.39%	1.6560
24	2.000	1.105	0.730	62.53%	1.5993
25	2.083	1.126	0.771	64.55%	1.5491
26	2.167	1.146	0.811	66.47%	1.5045
27	2.250	1.165	0.848	68.28%	1.4646
28	2.333	1.183	0.883	69.99%	1.4287
29	2.417	1.201	0.917	71.62%	1.3963
30	2.500	1.217	0.948	73.15%	1.3671
31	2.583	1.232	0.978	74.60%	1.3405
32	2.667	1.246	1.006	75.97%	1.3163
33	2.750	1.260	1.033	77.27%	1.2941
34	2.833	1.273	1.058	78.50%	1.2739
35	2.917	1.285	1.082	79.66%	1.2553
36	3.000	1.297	1.105	80.76%	1.2382

Exhibit 1B Clark-Robbin Two Factor Approximation of AY Development

Two Factor Formula: $F_{A+T}(t) \approx F_A(t)*F_T(r_A(t))$

Formula with Error Term: $F_{A+T}(t) = F_A(t)*F_T(r_A(t))+ 1/2*F_A(t)*v_A(t)*f_T'(r_A(t))$

		A = AY				Two-Factor AMOL Approx			Formula with Error Term		
		$F_A(t)$	$m_A(t)$	$v_A(t)$	$r_A(t)$ = $t-m_A(t)$	$F_T(r_A(t))$	$F_A(t)$ * $F_T(r_A(t))$		$f'_T(r_A(t))$		
t months	t in years	ETD%	Avg Date of Loss (Years)	Variance of Date of Loss	Maturity of Loss (Years)	PCT of T at Condl AMOL	PCT ULT Approx	Error	Endpt eval	PCT ULT Approx	Error
0	0.000		0.000		0.000						
1	0.083	8.33%	0.042	0.001	0.042	2.74%	0.23%	0.00%	-0.432	0.23%	0.00%
2	0.167	16.67%	0.083	0.002	0.083	5.40%	0.90%	0.01%	-0.420	0.89%	0.00%
3	0.250	25.00%	0.125	0.005	0.125	8.00%	2.00%	0.03%	-0.409	1.97%	0.00%
4	0.333	33.33%	0.167	0.009	0.167	10.52%	3.51%	0.06%	-0.398	3.44%	0.00%
5	0.417	41.67%	0.208	0.014	0.208	12.97%	5.40%	0.12%	-0.387	5.29%	0.00%
6	0.500	50.00%	0.250	0.021	0.250	15.35%	7.68%	0.20%	-0.376	7.48%	0.00%
7	0.583	58.33%	0.292	0.028	0.292	17.67%	10.31%	0.30%	-0.366	10.01%	0.00%
8	0.667	66.67%	0.333	0.037	0.333	19.93%	13.28%	0.44%	-0.356	12.84%	0.00%
9	0.750	75.00%	0.375	0.047	0.375	22.12%	16.59%	0.61%	-0.346	15.98%	0.00%
10	0.833	83.33%	0.417	0.058	0.417	24.25%	20.21%	0.81%	-0.337	19.40%	0.00%
11	0.917	91.67%	0.458	0.070	0.458	26.33%	24.13%	1.06%	-0.327	23.08%	0.00%
12	1.000	100.00%	0.500	0.083	0.500	28.35%	28.35%	1.33%	-0.318	27.02%	0.01%
13	1.083	100.00%	0.500	0.083	0.583	32.22%	32.22%	1.26%	-0.301	30.96%	0.01%
14	1.167	100.00%	0.500	0.083	0.667	35.88%	35.88%	1.19%	-0.285	34.69%	0.01%
15	1.250	100.00%	0.500	0.083	0.750	39.35%	39.35%	1.13%	-0.270	38.22%	0.01%
16	1.333	100.00%	0.500	0.083	0.833	42.62%	42.62%	1.07%	-0.255	41.56%	0.01%
17	1.417	100.00%	0.500	0.083	0.917	45.73%	45.73%	1.01%	-0.241	44.72%	0.01%
18	1.500	100.00%	0.500	0.083	1.000	48.66%	48.66%	0.96%	-0.228	47.71%	0.01%
19	1.583	100.00%	0.500	0.083	1.083	51.43%	51.43%	0.90%	-0.216	50.53%	0.01%
20	1.667	100.00%	0.500	0.083	1.167	54.06%	54.06%	0.86%	-0.204	53.21%	0.00%
21	1.750	100.00%	0.500	0.083	1.250	56.54%	56.54%	0.81%	-0.193	55.74%	0.00%
22	1.833	100.00%	0.500	0.083	1.333	58.89%	58.89%	0.77%	-0.183	58.13%	0.00%
23	1.917	100.00%	0.500	0.083	1.417	61.11%	61.11%	0.72%	-0.173	60.39%	0.00%
24	2.000	100.00%	0.500	0.083	1.500	63.21%	63.21%	0.69%	-0.164	62.53%	0.00%
25	2.083	100.00%	0.500	0.083	1.583	65.20%	65.20%	0.65%	-0.155	64.56%	0.00%
26	2.167	100.00%	0.500	0.083	1.667	67.08%	67.08%	0.61%	-0.146	66.47%	0.00%
27	2.250	100.00%	0.500	0.083	1.750	68.86%	68.86%	0.58%	-0.138	68.28%	0.00%
28	2.333	100.00%	0.500	0.083	1.833	70.54%	70.54%	0.55%	-0.131	70.00%	0.00%
29	2.417	100.00%	0.500	0.083	1.917	72.13%	72.13%	0.52%	-0.124	71.62%	0.00%
30	2.500	100.00%	0.500	0.083	2.000	73.64%	73.64%	0.49%	-0.117	73.15%	0.00%
31	2.583	100.00%	0.500	0.083	2.083	75.06%	75.06%	0.46%	-0.111	74.60%	0.00%
32	2.667	100.00%	0.500	0.083	2.167	76.41%	76.41%	0.44%	-0.105	75.98%	0.00%
33	2.750	100.00%	0.500	0.083	2.250	77.69%	77.69%	0.42%	-0.099	77.27%	0.00%
34	2.833	100.00%	0.500	0.083	2.333	78.89%	78.89%	0.39%	-0.094	78.50%	0.00%
35	2.917	100.00%	0.500	0.083	2.417	80.03%	80.03%	0.37%	-0.089	79.66%	0.00%
36	3.000	100.00%	0.500	0.083	2.500	81.11%	81.11%	0.35%	-0.084	80.76%	0.00%

Exhibit 2 **AY Average Date of Loss Approximation of Policy Year**

			AY Standard ADOL Approximation of PY						
		A = AY	A = AY m=6 months B = PY m = 12 months Δ = 6 months						
		$F_{A+T}(t)$	t-Δ months	$F_{A+T}(t-\Delta)$	$F_{B+T}(t)$				
t months	t in years	AY PCT ULT	Shifted eval	ADOL Approx PCT ULT	True PY PCT ULT	Error	ADOL Approx ATU LDF	True PY ATU LDF	Error
0	0.000	0.00%	-6	N/A	0.00%	N/A			
1	0.083	0.23%	-5	N/A	0.01%	N/A	N/A	15,768.597	N/A
2	0.167	0.89%	-4	N/A	0.05%	N/A	N/A	1,998.297	N/A
3	0.250	1.97%	-3	N/A	0.17%	N/A	N/A	600.197	N/A
4	0.333	3.44%	-2	N/A	0.39%	N/A	N/A	256.647	N/A
5	0.417	5.29%	-1	N/A	0.75%	N/A	N/A	133.173	N/A
6	0.500	7.48%	0	0.00%	1.28%	-1.28%	N/A	78.097	N/A
7	0.583	10.00%	1	2.73%	2.01%	0.72%	36.670	49.832	(13.162)
8	0.667	12.84%	2	5.36%	2.96%	2.40%	18.673	33.822	(15.149)
9	0.750	15.98%	3	7.89%	4.16%	3.73%	12.676	24.064	(11.388)
10	0.833	19.40%	4	10.33%	5.63%	4.70%	9.679	17.769	(8.090)
11	0.917	23.08%	5	12.69%	7.40%	5.29%	7.882	13.521	(5.640)
12	1.000	27.01%	6	14.96%	9.48%	5.48%	6.685	10.547	(3.862)
13	1.083	30.96%	7	17.15%	11.89%	5.26%	5.831	8.409	(2.579)
14	1.167	34.69%	8	19.27%	14.58%	4.68%	5.191	6.857	(1.666)
15	1.250	38.22%	9	21.31%	17.51%	3.80%	4.693	5.712	(1.018)
16	1.333	41.56%	10	23.28%	20.61%	2.67%	4.296	4.852	(0.556)
17	1.417	44.71%	11	25.18%	23.84%	1.33%	3.972	4.194	(0.222)
18	1.500	47.70%	12	27.01%	27.17%	-0.15%	3.702	3.681	0.021
19	1.583	50.53%	13	30.96%	30.53%	0.42%	3.230	3.275	(0.045)
20	1.667	53.20%	14	34.69%	33.91%	0.78%	2.883	2.949	(0.066)
21	1.750	55.73%	15	38.22%	37.25%	0.97%	2.617	2.685	(0.068)
22	1.833	58.12%	16	41.56%	40.52%	1.04%	2.406	2.468	(0.061)
23	1.917	60.39%	17	44.71%	43.69%	1.02%	2.236	2.289	(0.052)
24	2.000	62.53%	18	47.70%	46.73%	0.97%	2.096	2.140	(0.044)
25	2.083	64.55%	19	50.53%	49.61%	0.92%	1.979	2.016	(0.037)
26	2.167	66.47%	20	53.20%	52.33%	0.87%	1.880	1.911	(0.031)
27	2.250	68.28%	21	55.73%	54.91%	0.82%	1.794	1.821	(0.027)
28	2.333	69.99%	22	58.12%	57.34%	0.78%	1.720	1.744	(0.023)
29	2.417	71.62%	23	60.39%	59.65%	0.74%	1.656	1.676	(0.020)
30	2.500	73.15%	24	62.53%	61.83%	0.70%	1.599	1.617	(0.018)
31	2.583	74.60%	25	64.55%	63.89%	0.66%	1.549	1.565	(0.016)
32	2.667	75.97%	26	66.47%	65.84%	0.62%	1.504	1.519	(0.014)
33	2.750	77.27%	27	68.28%	67.69%	0.59%	1.465	1.477	(0.013)
34	2.833	78.50%	28	69.99%	69.44%	0.56%	1.429	1.440	(0.011)
35	2.917	79.66%	29	71.62%	71.09%	0.53%	1.396	1.407	(0.010)
36	3.000	80.76%	30	73.15%	72.65%	0.50%	1.367	1.376	(0.009)

Exhibit 3A

AMOL PY Cut-off

AMOL Dates and Exposure Earning Adjustment

Policy Year cut-off

	B= UWY Cutoff			A = AY			AY Equivalent				
						$r_A(t)$ =			$r_A(t^*)$ =		
	$F_B(t)$	$m_B(t)$	t-$m_B(t)$	$F_A(t)$	$m_A(t)$	t-$m_A(t)$	t^*	$m_A(t^*)$	t^*-$m_A(t^*)$	$F_A(t^*)$	$F_B(t)/F_A(t^*)$
t months	ETD%	Condl Avg Date of Loss (Months)	Avg Maturity of Loss (Months)	ETD%	Condl Avg Date of Loss (Months)	Avg Maturity of Loss (Months)	Eval Age for AY Equiv Maturity (Months)	Condl Avg Date of Loss (Months)	Avg Maturity	ETD % at Eval Age	ETD % Adj Factor
0	0.00%	0.000	0.000		0.000	0.000					
1	0.69%	0.667	0.333	8.33%	0.500	0.500	0.667	0.333	0.333	5.56%	12.500%
2	2.78%	1.333	0.667	16.67%	1.000	1.000	1.333	0.667	0.667	11.11%	25.000%
3	6.25%	2.000	1.000	25.00%	1.500	1.500	2.000	1.000	1.000	16.67%	37.500%
4	11.11%	2.667	1.333	33.33%	2.000	2.000	2.667	1.333	1.333	22.22%	50.000%
5	17.36%	3.333	1.667	41.67%	2.500	2.500	3.333	1.667	1.667	27.78%	62.500%
6	25.00%	4.000	2.000	50.00%	3.000	3.000	4.000	2.000	2.000	33.33%	75.000%
7	34.03%	4.667	2.333	58.33%	3.500	3.500	4.667	2.333	2.333	38.89%	87.500%
8	44.44%	5.333	2.667	66.67%	4.000	4.000	5.333	2.667	2.667	44.44%	100.000%
9	56.25%	6.000	3.000	75.00%	4.500	4.500	6.000	3.000	3.000	50.00%	112.500%
10	69.44%	6.667	3.333	83.33%	5.000	5.000	6.667	3.333	3.333	55.56%	125.000%
11	84.03%	7.333	3.667	91.67%	5.500	5.500	7.333	3.667	3.667	61.11%	137.500%
12	100.00%	8.000	4.000	100.00%	6.000	6.000	8.000	4.000	4.000	66.67%	150.000%
13	100.00%	8.000	5.000	100.00%	6.000	7.000	10.000	5.000	5.000	83.33%	120.000%
14	100.00%	8.000	6.000	100.00%	6.000	8.000	12.000	6.000	6.000	100.00%	100.000%
15	100.00%	8.000	7.000	100.00%	6.000	9.000	13.000	6.000	7.000	100.00%	100.000%
16	100.00%	8.000	8.000	100.00%	6.000	10.000	14.000	6.000	8.000	100.00%	100.000%
17	100.00%	8.000	9.000	100.00%	6.000	11.000	15.000	6.000	9.000	100.00%	100.000%
18	100.00%	8.000	10.000	100.00%	6.000	12.000	16.000	6.000	10.000	100.00%	100.000%
19	100.00%	8.000	11.000	100.00%	6.000	13.000	17.000	6.000	11.000	100.00%	100.000%
20	100.00%	8.000	12.000	100.00%	6.000	14.000	18.000	6.000	12.000	100.00%	100.000%
21	100.00%	8.000	13.000	100.00%	6.000	15.000	19.000	6.000	13.000	100.00%	100.000%
22	100.00%	8.000	14.000	100.00%	6.000	16.000	20.000	6.000	14.000	100.00%	100.000%
23	100.00%	8.000	15.000	100.00%	6.000	17.000	21.000	6.000	15.000	100.00%	100.000%
24	100.00%	8.000	16.000	100.00%	6.000	18.000	22.000	6.000	16.000	100.00%	100.000%
25	100.00%	8.000	17.000	100.00%	6.000	19.000	23.000	6.000	17.000	100.00%	100.000%
26	100.00%	8.000	18.000	100.00%	6.000	20.000	24.000	6.000	18.000	100.00%	100.000%
27	100.00%	8.000	19.000	100.00%	6.000	21.000	25.000	6.000	19.000	100.00%	100.000%
28	100.00%	8.000	20.000	100.00%	6.000	22.000	26.000	6.000	20.000	100.00%	100.000%
29	100.00%	8.000	21.000	100.00%	6.000	23.000	27.000	6.000	21.000	100.00%	100.000%
30	100.00%	8.000	22.000	100.00%	6.000	24.000	28.000	6.000	22.000	100.00%	100.000%
31	100.00%	8.000	23.000	100.00%	6.000	25.000	29.000	6.000	23.000	100.00%	100.000%
32	100.00%	8.000	24.000	100.00%	6.000	26.000	30.000	6.000	24.000	100.00%	100.000%
33	100.00%	8.000	25.000	100.00%	6.000	27.000	31.000	6.000	25.000	100.00%	100.000%
34	100.00%	8.000	26.000	100.00%	6.000	28.000	32.000	6.000	26.000	100.00%	100.000%
35	100.00%	8.000	27.000	100.00%	6.000	29.000	33.000	6.000	27.000	100.00%	100.000%
36	100.00%	8.000	28.000	100.00%	6.000	30.000	34.000	6.000	28.000	100.00%	100.000%

Exhibit 3B **AMOL PY Cut-off**

Average Maturity Approximation and Error Comparision

Policy Year cut-off

t	$F_{T\|A}(t)$	t*	$F_{T\|A}(t^*)$	$F_B(t)/F_A(t^*)$	$F^*_{T\|B}(t)$	$F_{T\|B}(t)$	$F^*_{T\|B}(t)$ - FT|B(t)	$F^*_{T\|B}(t)$	$F_{T\|B}(t)$	$F^*_{T\|B}(t)$ - FT|B(t)
t in months	AY PCT of ULT	Equivalent Eval Age	AY PCT ULT at Equivalent Eval Age	ETD % Adj Factor	AMOL Approx of PY Cutoff PCT ULT	Exact PY Cutoff PCT ULT	Error	AMOL Approx of PY Cutoff ATU LDF	Exact PY Cutoff ATU LDF	Error
0	0.000%					0.000%	0.000%		-	-
1	0.227%	0.667	0.152%	12.500%	0.019%	0.013%	0.006%	#####	#####	#####
2	0.893%	1.333	0.449%	25.000%	0.112%	0.100%	0.012%	890.8213	999.1486	(108.3273)
3	1.972%	2.000	0.893%	37.500%	0.335%	0.333%	0.001%	298.7647	300.0986	(1.3339)
4	3.444%	2.667	1.612%	50.000%	0.806%	0.779%	0.027%	124.0417	128.3236	(4.2819)
5	5.286%	3.333	2.463%	62.500%	1.539%	1.502%	0.037%	64.9662	66.5866	(1.6204)
6	7.480%	4.000	3.444%	75.000%	2.583%	2.561%	0.022%	38.7153	39.0486	(0.3333)
7	10.005%	4.667	4.672%	87.500%	4.088%	4.013%	0.075%	24.4604	24.9161	(0.4557)
8	12.844%	5.333	6.018%	100.000%	6.018%	5.913%	0.104%	16.6181	16.9111	(0.2930)
9	15.980%	6.000	7.480%	112.500%	8.415%	8.311%	0.103%	11.8840	12.0320	(0.1479)
10	19.396%	6.667	9.163%	125.000%	11.454%	11.255%	0.198%	8.7307	8.8846	(0.1539)
11	23.079%	7.333	10.951%	137.500%	15.058%	14.791%	0.266%	6.6411	6.7607	(0.1196)
12	27.013%	8.000	12.844%	150.000%	19.266%	18.962%	0.303%	5.1906	5.2736	(0.0830)
13	30.957%	10.000	19.396%	120.000%	23.276%	23.342%	-0.066%	4.2963	4.2842	0.0121
14	34.688%	12.000	27.013%	100.000%	27.013%	27.484%	-0.472%	3.7020	3.6384	0.0635
15	38.217%	13.000	30.957%	100.000%	30.957%	31.403%	-0.446%	3.2303	3.1844	0.0459
16	41.556%	14.000	34.688%	100.000%	34.688%	35.110%	-0.422%	2.8828	2.8482	0.0347
17	44.715%	15.000	38.217%	100.000%	38.217%	38.617%	-0.399%	2.6166	2.5895	0.0271
18	47.702%	16.000	41.556%	100.000%	41.556%	41.934%	-0.378%	2.4064	2.3847	0.0217
19	50.528%	17.000	44.715%	100.000%	44.715%	45.072%	-0.357%	2.2364	2.2187	0.0177
20	53.202%	18.000	47.702%	100.000%	47.702%	48.040%	-0.338%	2.0963	2.0816	0.0147
21	55.731%	19.000	50.528%	100.000%	50.528%	50.848%	-0.320%	1.9791	1.9666	0.0124
22	58.123%	20.000	53.202%	100.000%	53.202%	53.504%	-0.302%	1.8796	1.8690	0.0106
23	60.386%	21.000	55.731%	100.000%	55.731%	56.017%	-0.286%	1.7943	1.7852	0.0092
24	62.527%	22.000	58.123%	100.000%	58.123%	58.394%	-0.271%	1.7205	1.7125	0.0080
25	64.552%	23.000	60.386%	100.000%	60.386%	60.642%	-0.256%	1.6560	1.6490	0.0070
26	66.468%	24.000	62.527%	100.000%	62.527%	62.769%	-0.242%	1.5993	1.5931	0.0062
27	68.280%	25.000	64.552%	100.000%	64.552%	64.781%	-0.229%	1.5491	1.5437	0.0055
28	69.994%	26.000	66.468%	100.000%	66.468%	66.684%	-0.217%	1.5045	1.4996	0.0049
29	71.616%	27.000	68.280%	100.000%	68.280%	68.485%	-0.205%	1.4646	1.4602	0.0044
30	73.149%	28.000	69.994%	100.000%	69.994%	70.188%	-0.194%	1.4287	1.4247	0.0039
31	74.600%	29.000	71.616%	100.000%	71.616%	71.799%	-0.183%	1.3963	1.3928	0.0036
32	75.973%	30.000	73.149%	100.000%	73.149%	73.323%	-0.174%	1.3671	1.3638	0.0032
33	77.271%	31.000	74.600%	100.000%	74.600%	74.765%	-0.164%	1.3405	1.3375	0.0029
34	78.500%	32.000	75.973%	100.000%	75.973%	76.128%	-0.155%	1.3163	1.3136	0.0027
35	79.662%	33.000	77.271%	100.000%	77.271%	77.418%	-0.147%	1.2941	1.2917	0.0025
36	80.761%	34.000	78.500%	100.000%	78.500%	78.639%	-0.139%	1.2739	1.2716	0.0023

Exhibit 4A

AMOL PY

AMOL Dates and Exposure Earning Adjustment

Policy Year

		B=UWY(PY)			A = AY			AY Equivalent				
		$F_B(t)$	$m_B(t)$	$t\text{-}m_B(t)$	$F_A(t)$	$m_A(t)$	$r_A(t) =$ $t\text{-}m_A(t)$	t^*	$m_A(t^*)$	$r_A(t^*) =$ $t^*\text{-}m_A(t^*)$	$F_A(t^*)$	$F_B(t)/F_A(t^*)$
t months	t in years	ETD%	Condl Avg Date of Loss (Months)	Avg Maturity of Loss (Months)	ETD%	Condl Avg Date of Loss (Months)	Avg Maturity of Loss (Months)	Eval Age for AY Equiv Maturity (Months)	Condl Avg Date of Loss (Months)	Avg Maturity	ETD % at AMOL Equiv Eval Age	ETD % Adj Factor
0	0.000	0.00%	0.000	0.000		0.000	0.000					
1	0.083	0.35%	0.667	0.333	8.33%	0.500	0.500	0.667	0.333	0.333	5.56%	6.250%
2	0.167	1.39%	1.333	0.667	16.67%	1.000	1.000	1.333	0.667	0.667	11.11%	12.500%
3	0.250	3.13%	2.000	1.000	25.00%	1.500	1.500	2.000	1.000	1.000	16.67%	18.750%
4	0.333	5.56%	2.667	1.333	33.33%	2.000	2.000	2.667	1.333	1.333	22.22%	25.000%
5	0.417	8.68%	3.333	1.667	41.67%	2.500	2.500	3.333	1.667	1.667	27.78%	31.250%
6	0.500	12.50%	4.000	2.000	50.00%	3.000	3.000	4.000	2.000	2.000	33.33%	37.500%
7	0.583	17.01%	4.667	2.333	58.33%	3.500	3.500	4.667	2.333	2.333	38.89%	43.750%
8	0.667	22.22%	5.333	2.667	66.67%	4.000	4.000	5.333	2.667	2.667	44.44%	50.000%
9	0.750	28.13%	6.000	3.000	75.00%	4.500	4.500	6.000	3.000	3.000	50.00%	56.250%
10	0.833	34.72%	6.667	3.333	83.33%	5.000	5.000	6.667	3.333	3.333	55.56%	62.500%
11	0.917	42.01%	7.333	3.667	91.67%	5.500	5.500	7.333	3.667	3.667	61.11%	68.750%
12	1.000	50.00%	8.000	4.000	100.00%	6.000	6.000	8.000	4.000	4.000	66.67%	75.000%
13	1.083	57.99%	8.619	4.381	100.00%	6.000	7.000	8.762	4.381	4.381	73.02%	79.411%
14	1.167	65.28%	9.163	4.837	100.00%	6.000	8.000	9.674	4.837	4.837	80.61%	80.975%
15	1.250	71.88%	9.652	5.348	100.00%	6.000	9.000	10.696	5.348	5.348	89.13%	80.640%
16	1.333	77.78%	10.095	5.905	100.00%	6.000	10.000	11.810	5.905	5.905	98.41%	79.032%
17	1.417	82.99%	10.497	6.503	100.00%	6.000	11.000	12.503	6.000	6.503	100.00%	82.986%
18	1.500	87.50%	10.857	7.143	100.00%	6.000	12.000	13.143	6.000	7.143	100.00%	87.500%
19	1.583	91.32%	11.176	7.824	100.00%	6.000	13.000	13.824	6.000	7.824	100.00%	91.319%
20	1.667	94.44%	11.451	8.549	100.00%	6.000	14.000	14.549	6.000	8.549	100.00%	94.444%
21	1.750	96.88%	11.677	9.323	100.00%	6.000	15.000	15.323	6.000	9.323	100.00%	96.875%
22	1.833	98.61%	11.850	10.150	100.00%	6.000	16.000	16.150	6.000	10.150	100.00%	98.611%
23	1.917	99.65%	11.961	11.039	100.00%	6.000	17.000	17.039	6.000	11.039	100.00%	99.653%
24	2.000	100.00%	12.000	12.000	100.00%	6.000	18.000	18.000	6.000	12.000	100.00%	100.000%
25	2.083	100.00%	12.000	13.000	100.00%	6.000	19.000	19.000	6.000	13.000	100.00%	100.000%
26	2.167	100.00%	12.000	14.000	100.00%	6.000	20.000	20.000	6.000	14.000	100.00%	100.000%
27	2.250	100.00%	12.000	15.000	100.00%	6.000	21.000	21.000	6.000	15.000	100.00%	100.000%
28	2.333	100.00%	12.000	16.000	100.00%	6.000	22.000	22.000	6.000	16.000	100.00%	100.000%
29	2.417	100.00%	12.000	17.000	100.00%	6.000	23.000	23.000	6.000	17.000	100.00%	100.000%
30	2.500	100.00%	12.000	18.000	100.00%	6.000	24.000	24.000	6.000	18.000	100.00%	100.000%
31	2.583	100.00%	12.000	19.000	100.00%	6.000	25.000	25.000	6.000	19.000	100.00%	100.000%
32	2.667	100.00%	12.000	20.000	100.00%	6.000	26.000	26.000	6.000	20.000	100.00%	100.000%
33	2.750	100.00%	12.000	21.000	100.00%	6.000	27.000	27.000	6.000	21.000	100.00%	100.000%
34	2.833	100.00%	12.000	22.000	100.00%	6.000	28.000	28.000	6.000	22.000	100.00%	100.000%
35	2.917	100.00%	12.000	23.000	100.00%	6.000	29.000	29.000	6.000	23.000	100.00%	100.000%
36	3.000	100.00%	12.000	24.000	100.00%	6.000	30.000	30.000	6.000	24.000	100.00%	100.000%

Exhibit 4B AMOL PY

Average Maturity Approximation and Error Comparision

Policy Year cut-off

					PCT ULT Approx vs Exact			ATU LDF Approx vs Exact		
t	$F_{T\|A}(t)$	t*	$F_{T\|A}(t^*)$	$F_B(t)/F_A(t^*)$	$F^*_{T\|B}(t)$	$F_{T\|B}(t)$	$F^*_{T\|B}(t)$ - FT|B(t)	$F^*_{T\|B}(t)$	$F_{T\|B}(t)$	$F^*_{T\|B}(t)$ - FT|B(t)
t in months	AY Loss PCT of ULT	Equivalent Eval Age	AY Loss PCT ULT at Equivalent Eval Age	ETD % Adj Factor	Approx UW PCT ULT	UWY PCT ULT exact	Error	AMOL Approx of PY ATU LDF	Exact PY ATU LDF	Error
0	0.000%					0.000%	0.000%		-	-
1	0.227%	0.667	0.102%	6.250%	0.006%	0.006%	0.000%	######	15,768.5972	(24.0061)
2	0.893%	1.333	0.402%	12.500%	0.050%	0.050%	0.000%	1,992.2948	1,998.2972	(6.0024)
3	1.972%	2.000	0.893%	18.750%	0.167%	0.167%	0.001%	597.5294	600.1972	(2.6678)
4	3.444%	2.667	1.568%	25.000%	0.392%	0.390%	0.002%	255.1467	256.6472	(1.5006)
5	5.286%	3.333	2.420%	31.250%	0.756%	0.751%	0.005%	132.2130	133.1732	(0.9602)
6	7.480%	4.000	3.444%	37.500%	1.291%	1.280%	0.011%	77.4306	78.0972	(0.6666)
7	10.005%	4.667	4.632%	43.750%	2.027%	2.007%	0.020%	49.3426	49.8322	(0.4896)
8	12.844%	5.333	5.980%	50.000%	2.990%	2.957%	0.033%	33.4476	33.8222	(0.3747)
9	15.980%	6.000	7.480%	56.250%	4.207%	4.156%	0.052%	23.7680	24.0639	(0.2959)
10	19.396%	6.667	9.127%	62.500%	5.705%	5.628%	0.077%	17.5298	17.7692	(0.2395)
11	23.079%	7.333	10.917%	68.750%	7.505%	7.396%	0.110%	13.3236	13.5214	(0.1978)
12	27.013%	8.000	12.844%	75.000%	9.633%	9.481%	0.152%	10.3812	10.5472	(0.1660)
13	30.957%	8.762	15.209%	79.411%	12.077%	11.892%	0.186%	8.2800	8.4092	(0.1292)
14	34.688%	9.674	18.252%	80.975%	14.779%	14.585%	0.195%	6.7662	6.8565	(0.0903)
15	38.217%	10.696	21.931%	80.640%	17.685%	17.507%	0.178%	5.6545	5.7119	(0.0574)
16	41.556%	11.810	26.245%	79.032%	20.742%	20.609%	0.132%	4.8212	4.8522	(0.0310)
17	44.715%	12.503	29.026%	82.986%	24.087%	23.844%	0.244%	4.1515	4.1939	(0.0424)
18	47.702%	13.143	31.503%	87.500%	27.565%	27.166%	0.399%	3.6278	3.6810	(0.0532)
19	50.528%	13.824	34.046%	91.319%	31.090%	30.534%	0.556%	3.2164	3.2750	(0.0586)
20	53.202%	14.549	36.650%	94.444%	34.614%	33.907%	0.707%	2.8890	2.9492	(0.0602)
21	55.731%	15.323	39.315%	96.875%	38.086%	37.248%	0.838%	2.6256	2.6847	(0.0591)
22	58.123%	16.150	42.042%	98.611%	41.458%	40.521%	0.937%	2.4121	2.4679	(0.0558)
23	60.386%	17.039	44.836%	99.653%	44.680%	43.692%	0.988%	2.2381	2.2888	(0.0506)
24	62.527%	18.000	47.702%	100.000%	47.702%	46.728%	0.974%	2.0963	2.1400	(0.0437)
25	64.552%	19.000	50.528%	100.000%	50.528%	49.607%	0.921%	1.9791	2.0158	(0.0368)
26	66.468%	20.000	53.202%	100.000%	53.202%	52.330%	0.871%	1.8796	1.9109	(0.0313)
27	68.280%	21.000	55.731%	100.000%	55.731%	54.907%	0.824%	1.7943	1.8213	(0.0269)
28	69.994%	22.000	58.123%	100.000%	58.123%	57.343%	0.780%	1.7205	1.7439	(0.0234)
29	71.616%	23.000	60.386%	100.000%	60.386%	59.649%	0.738%	1.6560	1.6765	(0.0205)
30	73.149%	24.000	62.527%	100.000%	62.527%	61.829%	0.698%	1.5993	1.6174	(0.0180)
31	74.600%	25.000	64.552%	100.000%	64.552%	63.892%	0.660%	1.5491	1.5651	(0.0160)
32	75.973%	26.000	66.468%	100.000%	66.468%	65.843%	0.624%	1.5045	1.5188	(0.0143)
33	77.271%	27.000	68.280%	100.000%	68.280%	67.689%	0.591%	1.4646	1.4773	(0.0128)
34	78.500%	28.000	69.994%	100.000%	69.994%	69.435%	0.559%	1.4287	1.4402	(0.0115)
35	79.662%	29.000	71.616%	100.000%	71.616%	71.087%	0.529%	1.3963	1.4067	(0.0104)
36	80.761%	30.000	73.149%	100.000%	73.149%	72.649%	0.500%	1.3671	1.3765	(0.0094)

APPENDIX

FORMULAS FOR ACCIDENT YEAR, POLICY YEAR, AND POLICY YEAR CUT-OFF EXPOSURE STATISTICS

Statistic	Accident Year	(A1)
Density	$f_A(t) = \left\{\begin{matrix} 1\, for\ 0<t<1 \\ 0\ otherwise \end{matrix}\right\}$	
CDF	$F_A(t) = \left\{\begin{matrix} t \quad for\ 0<t<1 \\ 1 \quad for\ t\geq 1 \end{matrix}\right\}$	
Average Date of Loss	$m_A(t) = \left\{\begin{matrix} \frac{1}{2}\cdot t\ \ for\ t<1 \\ \frac{1}{2}\ \ for\ t\geq 1 \end{matrix}\right\}$	
Average Maturity of Loss	$r_A(t) = \left\{\begin{matrix} \frac{1}{2}\cdot t\ \ for\ t<1 \\ t-\frac{1}{2}\ for\ t\geq 1 \end{matrix}\right\}$	
Variance of Loss Exposure Date	$v_A(t) = \left\{\begin{matrix} \frac{1}{12}\cdot t^2\ \ for\ t<1 \\ \frac{1}{12}\ \ for\ t\geq 1 \end{matrix}\right\}$	

	Cut-off Policy Year	(A2)
Density	$f_A(t) = \left\{\begin{matrix} 2t\ \ for\ 0<t<1 \\ 0\ otherwise \end{matrix}\right\}$	
CDF	$F_A(t) = \left\{\begin{matrix} t^{\,2}\ \ for\ t<1 \\ 1\ \ for\ t\geq 1 \end{matrix}\right\}$	
Average Date of Loss	$m_A(t) = \left\{\begin{matrix} \frac{2}{3}\cdot t\ \ for\ t<1 \\ \frac{2}{3}\ \ for\ t\geq 1 \end{matrix}\right\}$	
Average Maturity of Loss	$r_A(t) = \left\{\begin{matrix} \frac{1}{3}\cdot t\ \ for\ t<1 \\ t-\frac{2}{3}\ for\ t\geq 1 \end{matrix}\right\}$	
Variance of Loss Exposure Date	$v_A(t) = \left\{\begin{matrix} \frac{1}{18}\cdot t^2\ \ for\ t<1 \\ \frac{1}{18}\ \ for\ t\geq 1 \end{matrix}\right\}$	

	Policy Year	(A3)

Density	$f_A(t) = \left\{\begin{matrix} t \ for\ 0<t<1 \\ 2-t\ for\ 1<t<2 \\ 0\ otherwise \end{matrix}\right\}$	
CDF	$F_A(t) = \left\{\begin{matrix} \frac{t^2}{2}\ for\ 0<t<1 \\ \frac{1}{2}+\frac{1}{2}(1-(2-t)^2)\ for\ 1<t<2 \\ 1\ \ for\ t>2 \end{matrix}\right\}$	
Average Date of Loss	$m_A(t) = \left\{\begin{matrix} \frac{2}{3}\cdot t\ \ for\ 0<t<1 \\ \frac{2}{3}+\frac{1}{3}\cdot\frac{(1-t\cdot(2-t)^2)}{F_A(t)}\ for\ 1<t<2 \\ 1\ for\ t>2 \end{matrix}\right\}$	
Average Maturity of Loss	$r_A(t) = \left\{\begin{matrix} \frac{1}{3}\cdot t\ \ for\ 0<t<1 \\ t-\frac{2}{3}-\frac{1}{3}\cdot\frac{(1-t\cdot(2-t)^2)}{F_A(t)} for\ 1<t<2 \\ t-1\ for\ t>2 \end{matrix}\right\}$	
Variance of Loss Exposure Date	$v_A(t) = \left\{\begin{matrix} \frac{1}{18}\cdot t^2\ \ for\ t<1 \\ \frac{1}{12}\cdot\frac{-2+8t^3-3t^4}{F_A(t)} - m_A(t)^2\ for\ \ 1<t<2 \\ \frac{1}{6}\ \ for\ t>2 \end{matrix}\right\}$	

REFERENCES

[1] Joseph Boor, "Interpolation Along a Curve", *Variance,* **Vol 8, Issue 1,** 2016, p.9-21.

[2] David Clark, "LDF Curve-Fitting and Stochastic Reserving: A Maximum Likelihood Approach", CAS Forum, **Fall 2006**, p 41-92.

[3] Ira Robbin, "Exposure Dependent Modeling of Percent of Ultimate Curves", *CAS Forum*, **Spring 2004**, p. 401-458.

[4] Ira Robbin and David Homer, "Analysis of Loss Development Patterns Using Infinitely Decomposable Percent of Ultimate Curves", *CAS Discussion Paper Program*, **May, 1988**, p501-538.

Abbreviations and Notations

ATA, Age-to-Age
ATU, Age-to-Ultimate
ETD, Earned to Date
LDF, Loss Development Factor
PCT ULT, Percent of Ultimate

Biography of the Author

Ira Robbin is currently Assistant Vice-President in Economic Capital Modeling at TransAtlantic Reinsurance in New York City. Ira received a Bachelors Degree in Math from Michigan State University and a PhD in Math from Rutgers University. He has served in a variety of research, actuarial pricing, reserving, and corporate roles over his career at companies including the Insurance Company of North America (INA), CIGNA Property and Casualty, ACE, Partner RE, Endurance, and AIG. While developing new techniques and theories, he has headed large risk property and casualty pricing units, developed pricing algorithms, constructed LDF fitting algorithms, produced price monitors, conducted reserve reviews, priced treaties, allocated capital, and computed ROE. He has written several Proceedings, Forum, and Study Note papers on a range of subjects, taught exam preparation classes and made numerous presentations at actuarial meetings.

Disclaimers

Opinions expressed in this paper are solely those of the author. They are not presented as the express or implied positions of the authors' current or prior employers or clients. No warranty is given that any formula or assertion is accurate. No liability is assumed whatsoever for any losses, direct or indirect, that may result from use of the methods described in this paper or reliance on any of the views expressed therein.

Loss Reserve Simulation Revisited

Richard L. Vaughan, FCAS, FSA, MAAA

Abstract. James Stanard's 1985 PCAS paper *A Simulation Test of Prediction Errors of Loss Reserve Estimation Techniques* was noteworthy as much for the simple and parsimonious estimators it introduced as for the simulation technique it used to test them. In subsequent years, those estimators have been widely adopted, a comprehensive loss development simulation model has been introduced by a working party of the CAS, and numerous new models of the loss process, with associated estimators, have been published. But there has been little further effort to compare such estimators by simulation in the manner of Stanard. In this paper we revisit the use of simulation to evaluate reserve estimators, and apply it to several recent models. We find good reason, as did Stanard, to prefer parsimonious conventional models, such as Bornhuetter-Ferguson with Cape Cod ELR's, to the chain-ladder model, and we find no evidence that any of the more recent estimators tested, most of which are elaborations of chain-ladder, performs any better.

1. INTRODUCTION

1.1 The use of simulation to test loss reserve estimators

James Stanard's 1985 PCAS paper *A Simulation Test of Prediction Errors of Loss Reserve Estimation Techniques* [10] was noteworthy in several respects. First, it demonstrated the use of simulation as a practical technique for evaluating the bias and efficiency of various reserving methods, using computers readily available to actuaries of the time. Second, it applied this simulation to test, among others, two simple and parsimonious techniques not well known before that date: Cape Cod and "additive". Third, it conclusively established the superiority of these new models over the conventional chain-ladder for the particular simulated data sets studied, and strongly suggested that this superiority would hold with more sophisticated simulations or indeed with real-world data.

In fact, Stanard's *results* have probably had a greater impact on actuarial practice than his *method*. For example, the Cape Cod (or Stanard-Buehlmann) estimator of expected loss ratios is now ubiquitous, and actuaries are universally aware of the pitfalls of chain-ladder projections in projecting recent accident years. But the further development of Stanard's simulation method has not been completely ignored. Some extensions were proposed by Vaughan in 1998 [11]. In 2011, the CAS Loss Development Simulation Working Party published an engine simulating the loss process, written in R. This simulator is in the public domain and available for download from the CAS web site [4], and it has stimulated additional research and development (see Shang [9]). But there have been few attempts to *apply* such tools to evaluate the performance of specific reserve estimators. This seems unfortunate in light of the many such estimators that have been proposed.

Here we illustrate how a small part of this void might be filled, by applying the CAS simulation model (actually a prototype of that model from 2007, written in APL) to evaluate several recent estimators and compare them with those originally evaluated by Stanard.

1.2 Simulation of the loss process

Simulation is the random generation of synthetic data, resembling what might be observed from a real-world process of interest. For reserving this is the *loss process*: the events surrounding the emergence of losses in a portfolio of insurance policies. These events include the occurrence and reporting of accidents, the true size of each reported loss, the initial and subsequent valuations of its case reserves, and the timing and amounts of payments and recoveries.

A simulation of the loss process starts from certain parameters that are not themselves simulated. These fixed meta-parameters permit us to customize each simulation to a particular line or lines of business, to a particular volume of exposure, to a particular company's claims-handling procedures, and to a particular economic and legal scenario. They include the exposures written in each time period, and parameters specifying the distributions of certain quantities, such as the frequency of accidents, the actual size of loss, the lag from accident to reporting, the lag from reporting to settlement, the size of any initial "fast-track" valuation, the waiting times between subsequent valuations, the case reserve error as a function of actual loss and time remaining until settlement, and so forth. The simulation proper uses random deviates from these distributions to generate many sample points, *each* of which is a *complete history of all transactions* resulting from all accidents covered by a hypothetical portfolio satisfying the starting assumptions, and each of which may be output in its full detail or after aggregating into triangles.

Some convenient forms for the distributions which model frequencies, waiting times, and dollar amounts are Poisson, negative binomial, exponential, Weibull, and lognormal. For verisimilitude it may be useful to go back one step further, and treat the parameters of these distributions, not as fixed, but as subject to random change, for example via a random walk to simulate the effects of turnover in the insured population. In this case the parameter in question may itself be given a distribution, the parameters of which become the new "fixed" ones.

A great many separate distributions may be involved. For example, we may wish to model multiple lines in a single simulation, or a line of business that includes several distinct types of loss potentially arising from a single accident, such as indemnity, medical, and expenses, or bodily injury and property damage, and we may wish frequencies to be correlated across lines or severity to be correlated with lag to payment. One of the advantages of simulation is that it may be as rich and complex an approximation to the real world as desired. When we are *estimating* we must avoid fitting too many parameters lest we obtain a good fit with no predictive value, but when we are *generating* simulated data there is no such constraint.

It goes without saying that a simulation intended for testing reserve estimators should extend to the final disposition of each claim incurred within the exposure period being studied, since we are interested in how well our estimators predict the entire "future" from the "known" part of the data.

The measures of bias and of variability that emerge from the simulation studies discussed here are global rather than conditional. They are averages across all simulated data sets whatever the "known" and "future" portion of each may be. At present it does not appear feasible to use simulation to study the distribution of "future" runoff conditional on a particular set of "known" data, such as a given combination of paid and incurred loss triangles, even though in principle this might be accomplished via Markov Chain Monte Carlo Bayesian analysis or otherwise. There are simply too many variables and too many cells involved.

1.3 Reserve estimators amenable to testing by simulation

It is the author's opinion that several of the most fruitful contributions to loss reserving methodology in recent decades have been among the simplest: the pure Bornhuetter-Ferguson (BF) method, with *a-priori* expected loss ratios (ELR's) [2], Buehlmann's BF method with Cape Cod estimate of ELR's [3], Stanard's "additive" method [10], and Spencer Gluck's 1997 enhancement of Cape Cod with decay factors [5]. Not only are these methods simple computationally, but they are parsimonious: they either *eliminate* the accident-year parameters of the conventional chain-ladder model (BF and additive), *reduce the number of* such parameters (Cape Cod), or *constrain their independence* (Gluck).

These simple enhancements reduce the need for judgment intervention. Judgment is troublesome and does not lend itself to simulation. First, it may be biased, especially when "tutored", i.e. when an estimate of the same reserve from a different source already exists, such as the carried reserve in an Actuarial Opinion situation. Second, there is no way of gauging the variability of reserves estimated by judgment. Third, it is not feasible to incorporate judgment when applying a reserve estimator, under program control, to each of many data sets.

On the other hand, some estimators do not need judgment intervention, or may be modified with protocols to achieve the goals of judgment automatically, a sort of "meta-judgment". The Bornhuetter-Ferguson method and its variations with Cape Cod and Gluck ELR's usually require little or no judgment, as does the additive estimator (which we shall henceforth call *Partial Loss Ratios*). But even these estimators may benefit from adjustment to the development-year parameters, especially at the later lags where experience is thin. In a one-off calculation the actuary may apply judgment to graduate the development or lag factors or to extend them with a tail. For a simulation, it is straightforward to automate such adjustments, either by blending experience with reference factors, by curve fitting, or by specifying the parameters of any necessary tail and executing them under program control. The underlying methods, together with such programmed adjustments, are very amenable to simulation studies.

The simulations described here generate complete transaction histories, so they may be used to test estimators based on individual claims. They also generate sets of matched triangles including paid and incurred losses and counts, so they may be used to test reserve estimators that rely on any one, or on more than one, of these triangles. In this paper we limit our attention to estimators based on triangles of various types.

2. BACKGROUND AND METHODS

2.1 The CAS Loss Development Simulator (CASLDS)

Starting in 2007, the CAS Loss Development Simulation Working Party, headed by Robert Bear and Mark Shapland, developed a software model, written in the language R, to simulate the loss process in considerable detail. This model was presented to the CAS in the *E-Forum* for Winter 2011 [4], and was extensively tested to confirm that the various distributions simulated had the expected statistical properties. The reader may download this model and its complete documentation from www.casact.org.

For this paper the author employed a prototype of CASLDS written in APL in 2007, and slightly improved since then, with which he was already familiar, because this prototype models the loss process in sufficient detail and saves its output to files that may be read conveniently by reserving programs written for this study. The author recommends the full R version of CASLDS, with its enhanced capabilities, for users wishing to do further work along these lines.

2.2 Simulation Procedure

There are two parts to our simulation procedure: (1) generating sets of loss histories, and (2) testing various reserve estimators by applying them to each loss history and comparing estimated reserves against "actual", i.e. simulated, runoff.

2.2.1. Loss histories

Each *set* of loss histories is a collection of 10000 *particular* loss histories developed with a common model and common parameters. We postulate (I) a "starter" model, with a single line of business, a single type of loss, no parameter drift, no trend, and a modest amount of case-reserve error (without which no model of incurred loss development would be realistic), (II) the addition of parameter drift; (III) the addition of trend, (IV) an increased amount of case reserve error, and (V) the addition of a second line of business, with lower frequency, greater severity, and greater trend, combined with the first for analysis. The parameters of these loss histories are shown in the following tables.

Loss history set	I	II	III	IV	V	
Description	Simple	Add drift	Add trend	Add wider case reserve errors	Complex	
Sample size	10000	10000	10000	10000	10000	
Accident years	10	10	10	10	10	
					Line 1	Line 2
Exposure	1	1	1	1	1	1
Lapse ratio (per month)	0	0.03	0.03	0.03	0.03	0.03
Development-month turnover	0	0.03	0.03	0.03	0.03	0.03
Frequency correlations across lines	1	1	1	1	1 0.25 0.25 1	
Frequency distribution	Poisson	Poisson	Poisson	Poisson	Poisson	Poisson
Frequency mean (per year)	50	50	50	50	50	25
Std dev of frequency means	0	8	8	8	8	5
Frequency trend by accident year	1	1	1.01	1.01	1.01	1.02
Report lag distribution	Weibull	Weibull	Weibull	Weibull	Weibull	Weibull
Report lag mean (days)	240	240	240	240	240	300
Report lag std dev	160	160	160	160	160	240
Std dev of report-lag means	0	24	24	24	24	40
Report lag minimum	0	0	0	0	0	0
Report lag maximum	1440	1440	1440	1440	1440	1600
Valuation lag distribution	Lognormal	Lognormal	Lognormal	Lognormal	Lognormal	Lognormal
Valuation lag mean (days)	90	90	90	90	90	90
Valuation lag std dev	30	30	30	30	30	30
Std dev of valuation-lag means	0	0	0	0	0	0
Valuation lag minimum	0	0	0	0	0	0
Valuation lag maximum	360	360	360	360	360	360
Payment lag distribution	Weibull	Weibull	Weibull	Weibull	Weibull	Weibull
Payment lag mean (days)	450	450	450	450	450	720
Payment lag std dev	300	300	300	300	300	480
Std dev of payment-lag means	0	40	40	40	40	80
Payment lag minimum	0	0	0	0	0	0
Payment lag maximum	2700	2700	2700	2700	2700	4500
Recovery lag distribution(*)	Expon.	Expon.	Expon.	Expon.	Expon.	Expon.
Recovery lag mean (days)	120	120	120	120	120	120
Std dev of recovery-lag means	0	0	0	0	0	0
Recovery lag minimum	0	0	0	0	0	0
Recovery lag maximum	730	730	730	730	730	730

(*) Recovery lag parameters are shown for completeness of the model. In the simulations performed here there are no recoveries, since the payment adequacy factors, shown below, are all 1.00.

Loss history set (continued)	I	II	III	IV	V	
Description	Simple	Add drift	Add trend	Add wider case reserve errors	Complex	
					Line 1	Line 2
Severity distribution	Lognormal	Lognormal	Lognormal	Lognormal	Lognormal	Lognormal
Severity mean	10000	10000	10000	10000	10000	15000
Severity std dev	30000	30000	30000	30000	30000	50000
Std dev of severity means	0	4000	4000	4000	4000	6000
Minimum severity	0	0	0	0	0	0
Maximum severity	1000000	1000000	1000000	1000000	1000000	1000000
Severity trend by accident year	1	1	1.03	1.03	1.03	1.04
Fraction of trend from acc to pmt	1	1	1	1	1	1
Correlation with payment lag	0	0	0	0	0	0.5
Probability spike at severity 0	0.1	0.1	0.1	0.1	0.1	0.1
Deductible	0	0	0	0	0	0
P(rounding to two signif digits)	0.2	0.2	0.2	0.2	0.2	0.2
Case reserve adequacy distribution	Lognormal	Lognormal	Lognormal	Lognormal	Lognormal	Lognormal
Mean adequacy at report date	1	1	1	1.10	1.10	1.20
Mean at 30% to settlement	1	1	1	1.10	1.10	1.15
Mean at 70% to settlement	1	1	1	1.05	1.05	1.05
Mean at 90% to settlement	1	1	1	0.95	0.95	0.95
Case reserve adequacy std dev	0.25	0.25	0.25	0.5	0.5	0.5
Std dev of adequacy means	0	0.1	0.1	0.1	0.1	0.1
P(adequacy=0 | actual=0)	0.6	0.6	0.6	0.6	0.6	0.6
P(adequacy=0 | actual>0)	0.05	0.05	0.05	0.05	0.05	0.05
Fast-track case reserve	4000	4000	4000	4000	4000	5000
P(rounding to two signif digits)	0.5	0.5	0.5	0.5	0.5	0.5
Inertia (weight to existing res)	0.2	0.2	0.2	0.2	0.2	0.2
Min material absolute change	100	100	100	100	100	100
Min material relative change	0.01	0.01	0.01	0.01	0.01	0.01
Payment adequacy factor distrib(*)	Lognormal	Lognormal	Lognormal	Lognormal	Lognormal	Lognormal
Mean payment adequacy factor	1.05	1.05	1.05	1.05	1.05	1.05
Payment adequacy factor std dev	0.1	0.1	0.1	0.1	0.1	0.1
Std dev of adequacy means	0	0	0	0	0	0
Probability spike at 1.00	1	1	1	1	1	1

(*) This section is included for completeness of the model. Because the probability spike at 1.00 is 1 for all simulations reported here, the payment adequacy factors have no effect.

A few explanations are in order.

The unit of exposure is arbitrary and here is simply taken to be a constant 1. Frequencies are defined relative to the initial total exposure. In this case there are no changes in exposure, over time, but, if there were any, the frequencies would be adjusted automatically to remain consistent.

One-parameter distributions are parametrized by their means and two-parameter distributions by their means and standard deviations; this makes the inputs similar for all distributions. The system converts these internally to the canonical parameters appropriate for generating random deviates from each distribution.

The means of most distributions may, if requested, be allowed to drift in the manner of a random walk, with the mean for each successive period equal to a weighted average of the existing mean and a mean drawn from a second-level Gamma distribution. The weight given to the newly-drawn mean is determined by a lapse rate or "turnover fraction". The concrete interpretation for parameters that change by accident period is that some inforce policies lapse and are replaced by new ones from a wider population. Parameters that change by development period lack this interpretation but we postulate a similar mechanism. Drift only applies to those parameters for which a non-zero "standard deviation of means" is specified in the foregoing tables.

Some quantities, while conveniently described over most of their range by a continuous distribution, also have nonzero probability spikes at certain values: for example, the probabilities that the actual loss will be zero, that the case reserve will be zero conditional on the actual loss's being nonzero, or that the case reserve will be zero conditional on the actual loss's being zero.

Recoveries are treated as the difference between the initial payment and the actual loss (net of deductible), depending on a distribution of initial-payment adequacy factors. This distribution in turn allows for a spike at 1.00; by setting the probability of this spike at 1, recoveries may be excluded from the simulation. This has been done here, so as to permit testing of reserve estimators requiring that the column sums of the incremental paid loss triangle be non-negative.

Case reserve adequacy is modeled by specifying a distribution with separate means at 0%, 30%, 70%, and 90% of the time from reporting to settlement (which is known to the simulation program even if it would not be known in real life). By setting these to values other than 1.00 we can model systematic bias in case reserves. Valuation lags may take any value, so case reserve adequacy at lags other than the above fractions of the settlement lag are handled by interpolation. A fast-track reserve overrides the sampled value at time of reporting, though not at subsequent valuations (including valuations interpolated between 0% and 30%). There is provision for "inertia" (the influence of the existing case reserve on the new value), for no change unless material, and for clustering at "round" numbers, as well as for spikes at zero.

Lag distributions are measured in days.

For most parameters, separate values may be specified for each month, and parameter drift also takes place month by month. Output may be in triangles of various cell sizes; here they are annual.

Certain parameters provided as options by the simulator but receiving the same value in all our simulations have been omitted from the above table; chief among these is provision for frequency seasonality factors, which are here, by default, all 1.

Our sets of loss histories as described above represent entirely imaginary lines of business. In practice the parameters should be selected, wherever possible, with reference to known characteristics of the real-world business for which the reserve estimators are to be used. Each set has a sample size of 10000 loss histories; this seems reasonable for our illustrations here, in light of the regularity of the results, but in practice the sample size may need to be increased for some lines, for example those with low frequencies and broad size-of-loss distributions. Each loss history covers a period of ten accident years; this also seems reasonable, in light of common practice.

Finally, the selected frequency distributions produce only a modest number of claims in each loss history, in most cases with a mean of 50 per year. This was a practical compromise in light of the fact that it is really the individual claims that are being simulated, each with its entire transaction history; most of the sets therefore contain the histories of some 5,000,000 claims, and the prototype simulator takes some time to generate these. All estimators produce a tighter distribution of results with larger triangles, and some require programmed "meta-judgment" adjustment for sparse triangles, but produce more regular results, without such adjustment, for larger triangles. To study this, we can sum the original 10,000 sample points by, say, fives, giving a new sample with only 2,000 points, but each with five times the original expected frequency.

2.2.2. Tests and comparisons

Our tests mainly address the question of which reserve estimators perform best against simulated loss histories of a particular type, but also may give some insight into which types of simulated loss history are best suited for evaluating particular reserve estimators.

We consider mature data, observed through a long enough development period to obtain direct estimates of lag factors to ultimate. For all five sets of loss histories described above, losses are essentially complete after at most 10 years, and the simulations run for ten accident years, so we use 10 x 10 lag triangles. Some of our simulations in fact push the final settlement of a few losses slightly beyond lag 10 years, but the amounts involved are immaterial.

Following Stanard, we calculate bias and root-mean-square error, each for the calculated reserve compared with the simulated runoff, and each expressed as a fraction relative to the runoff averaged across all loss histories. The simulated runoffs are identical across all tests and their averages across the samples are given in the following table:

Loss history set	I	II	III	IV	V
Runoff	849,729	847,105	1,292,773	1,294,504	3,789,979

Most of the estimators studied may be applied either to paid or incurred loss triangles. Some also use triangle of claim counts, closed or reported. Some use both paid and incurred information and produce a combined indication. In the tables below, we include a "combined" result for nearly all estimators. In most cases this is derived from a simple average of the paid and incurred results, sample point by sample point, but, as mentioned, some estimators combine the paid and incurred information in more sophisticated ways.

The estimators we have tested use:

1. BF with lag factors from LDF's and with ELR's determined using Gluck factors 1 (Cape Cod), 0 (straight chain-ladder), and various intermediate values.
2. BF with lag factors by Partial Loss Ratios (PLR), but with ELR's determined using Gluck factors 1 (equivalent to straight PLR, or Stanard's Additive method), 0 (for a "PLR chain-ladder" model determining each year's ELR from its own immature value), and various intermediate values.
3. Quarg & Mack, Munich Chain Ladder (MCL) [8].
4. Merz & Wütrich, Paid-Incurred Chain Reserving Method (PIC) [6] with non-informative priors.
5. Yamashiro, Recursive Credibility estimator [12].
6. Agbeko et al, Incurred Double Chain Ladder (DCL, IDCL, and BDCL) [1].
7. Müller, Affine Age-to-Age Development [7]: the models which Müller calls Generalized Linear Regression (ALDGLR; two parameters, constant weights) and Generalized Chain Ladder (ALDGCL; two parameters, weights equal to latest known losses).

When trend is included in the simulated loss histories, we run the foregoing estimators "under" trending, i.e. we assume the actuary has an a-priori accident-year trend factor, with which we trend all losses going into the calculations to a common reference date, and then perform the inverse process on the results.

Adjustments to published methods. Several of the published methods require adjustments to work properly with the small data sets generated by our simulation. Specifically:

1. The MCL method of Quarg and Mack develops the paid and incurred losses separately but uses correlations between P/I ratios and residuals of paid development factors, and between I/P ratios and residuals of incurred development factors, to bring the two estimates closer together. When applied to our sample data it is affected by the fact that for many sample triangles, development ends earlier than the last lag, so that the sample standard deviations of the later development factors are zero. Quarg and Mack suggest using judgment to reduce the size of the triangles in such cases, but we found it possible to work with the full 10 x 10 triangles in all cases.

Also, in their example, Quarg and Mack arbitrarily set the standard deviation of the development factors at the last lag (n.a. because of only a single data point) to 0.1; we replaced this with a value extrapolated as suggested in Mack (1993), and also used by Merz and Wüthrich (see below).

Finally, large ratios of σ to ρ can produce adjusted development factors (paid or incurred) that are unacceptably large or small, in some cases producing infinite or negative cumulative projected losses. To reduce this effect, as part of the rules for extrapolating the final value of the paid or incurred σ, we subject it to a maximum equal to the corresponding ρ, derived from the standard deviations of the I/P or P/I ratios.

2. The PIC method of Merz and Wüthrich produces three separate estimates, one conditional on the known paid triangle, one conditional on the known incurred triangle, and one conditional on both known triangles. While these estimates work well for large data sets, the method sometimes fails to produce results for data sets of the size studied here. Moreover, the estimates from paid data may be unusable because of extremely large projections of later accident years. This is the result of using unweighted estimates of σ, the sample standard deviation of log(paid development factors). Accordingly, we added an option of weighting the estimates with the denominator losses in the development factors, and used this option for our simulation runs; this greatly reduced bias and RMS error for the paid estimates and slightly reduced them for the other estimates.

   To avoid zero values in the vector of standard deviations, which cause the rest of the PIC estimator to fail, we set them equal to the nearest preceding nonzero value, if available, else the nearest following nonzero value. Similarly, to avoid a zero value in the southwest corner cell of the paid or incurred triangle, we replaced any such values with the average of all values at lag 12 months; this may introduce some positive bias to the results, but it would be a reasonable judgment adjustment in clinical practice.

   Even with these adjustments, the Merz & Wüthrich method fails for some data sets, usually because of difficulties inverting the matrices required for the incurred-only and the combined estimates. In the tables below, we show the number of samples for which results were available, and the bias and RMS errors across these samples only.

3. With some data sets, Yamashiro's Recursive Credibility estimator can project cumulative losses that are negative for one or more years. We treat such results as unavailable, and, in the tables below, we show the number of samples for which results were available, and the bias and RMS errors across these samples only. Another approach is to modify the data, by adding a small constant to each cell of each cumulative triangle entering the calculations, and subtracting it from the results, increasing the constant as necessary until any negative

cumulative projections disappear. While this eliminates the handful of failed estimates, its other properties are unknown.

4. The estimation of the report-to-payment lag factors (π-hat) for the DCL method as described by Agbeko et al involves the solution of a system of linear equations relating the estimated accident-to-report lags β-hat and the accident-to-payment lags β-tilde-hat. Agbeko et al recognize that this can produce negative results or results not summing to 1.00 and suggest a simple adjustment to correct, approximately, for this problem. We find that these estimates of π-hat, whether or not they include negative values, can lead to very erratic projections of ultimate losses, and we estimate π-hat directly from the triangle of paid losses by report date versus payment lag. The additional triangle that this requires is easily generated by the simulator.

5. Thomas Müller's Affine Loss Development requires a modest adjustment to avoid matrices that are not invertible. Specifically, when the weights matrix *W* has diagonal elements drawn from the losses known through lag *j*, it occasionally happens in our samples that one or more of these known losses is zero. To avoid this, we restrict *W* to its nonzero rows and columns, and we restrict the design matrix to its corresponding rows. Note also that the matrix inversions involved in this method should be performed with extended numerical precision, as some of the sample triangles lead to ill-conditioned matrices.

   In the Affine Loss Development models proper – those where the regression from development year to development year involves both a constant *c* and a factor *f* – Müller suggests reducing the model to a factor *f*, with no constant term, when projecting the last accident year, for which there is not enough data to estimate both *c* and *f*. We found that even for earlier development years it is possible for the estimation of *c* and *f* to produce erratic results, and that some adjustment is called for. It is tempting to try to create rules of "meta-judgment" to switch to a factor-only model depending on a preliminary calculation of *c* and *f*, but it is very difficult to devise such rules without introducing bias. Therefore we settled on applying the factor-only model automatically for the last two lags, rather than just the last one.

2.3 Simulation Results

2.3.1. Simple model.

The results of the simple model are summarized in the table below. BF stands for Bornhuetter-Ferguson, with lag factors determined by chain-ladder and with ELR's determined by Cape Cod with Gluck decay factor G. This family includes both the pure Cape Cod and the pure chain-ladder estimators. PLR BF stands for BF with lag factors determined by partial loss ratios and ELR's determined by Cape Cod with Gluck decay factor G; this family includes both the pure PLR and the PLR chain-ladder estimators. We defer discussion of Gluck factors intermediate between 0 and 1.

The "Combined" column contains the arithmetic mean of the paid and incurred values for most estimators (following an often-used practice), and contains an estimate explicitly conditional on both the paid and incurred known triangles in the case of Merz & Wüthrich.

In the following table the bias values for which the value of zero cannot be rejected with confidence 95% are printed in **bold** type; in some cases this may reflect the fact that the RMS error is large rather than that the absolute bias is small.

		Paid		Incurred		Combined	
Estimator	N	Bias	Error	Bias	Error	Bias	Error
BF, G=1 (pure Cape Cod)	10000	**0.00036**	0.39784	**0.00100**	0.26919	**0.00068**	0.31659
BF, G=0 (pure chain-ladder)	10000	0.23931	1.78908	0.01765	0.37349	0.12848	0.99645
PLR BF, G=1 (pure PLR)	10000	**-0.00289**	0.39601	**0.00015**	0.27094	**-0.00137**	0.31787
PLR BF, G=0 (PLR CL)	10000	**0.00502**	0.78747	**-0.00330**	0.32352	**0.00086**	0.50225
Quarg & Mack, MCL	10000	-0.06160	0.42254	0.03487	0.44314	-0.01337	0.39335
Merz & Wütrich, PIC	9993	0.12856	1.40853	0.07989	0.38262	0.08306	0.49093
Yamashiro, RC	9986	**-0.00742**	0.40880	**-0.00742**	0.40880	**-0.00742**	0.40880
Agbeko et al, DCL and IDCL	10000	0.22659	1.70207	-0.03438	0.40466	0.09611	0.96363
Agbeko et al, BDCL	10000	0.02515	0.42436				
Müller, ALD (GLR)	10000	**0.00079**	0.49288	0.00624	0.30889	**0.00352**	0.35347
Müller, ALD (GCL)	10000	**0.00141**	0.52837	0.00608	0.30941	**0.00374**	0.36729

The comparisons of chain ladder with Cape Cod and PLR confirm Stanard's findings, as expected. Even for this set of simple loss histories the pure chain ladder shows very high bias and RMS error.

The performance of pure Cape Cod and pure PLR are essentially indistinguishable. The incurred estimates outperform the paid, but this may be dependent on the arbitrary reserve error distributions assumed here.

MCL performs markedly better than the conventional chain-ladder, but its paid estimates have a fairly large negative bias, probably due to the fact that the most recent accident year, if zero after 12 months, will be projected to an ultimate of zero. This issue is a direct result of our small data sets. PIC outperforms the paid chain ladder but not the incurred. Yamashiro performs better than either paid or incurred chain ladder but not quite as well as pure Cape Cod or pure PLR. DCL performs marginally better than conventional chain ladder. The two Affine Loss Development models perform very well, despite requiring the estimation of twice as many parameters as BF or Chain Ladder to describe the development pattern.

2.3.2. Model with parameter drift

The results of the model with parameter drift added are summarized below.

		Paid		Incurred		Combined	
Estimator	N	Bias	Error	Bias	Error	Bias	Error
BF, G=1 (pure Cape Cod)	10000	**0.00391**	0.41527	**0.00088**	0.27587	**0.00239**	0.32967
BF, G=0 (pure chain-ladder)	10000	0.28084	2.04626	0.02077	0.38274	0.1508	1.12760
PLR BF, G=1 (pure PLR)	10000	**0.00127**	0.41661	**0.00021**	0.27912	**0.00074**	0.33322
PLR BF, G=0 (PLR CL)	10000	0.01618	0.81869	**-0.00202**	0.32334	**0.00708**	0.51726
Quarg & Mack, MCL	10000	-0.05888	0.45841	0.03437	0.45158	-0.01225	0.42037
Merz & Wütrich, PIC	9992	0.15817	1.62550	0.08237	0.38986	0.08586	0.48937
Yamashiro, RC	10000	**-0.00199**	0.48433	**-0.00199**	0.48433	**-0.00199**	0.48433
Agbeko et al, DCL and IDCL	10000	0.26652	1.94872	-0.02969	0.41510	0.11842	1.08768
Agbeko et al, BDCL	10000	0.02910	0.43282				
Müller, ALD (GLR)	10000	**-0.00229**	0.49000	**0.00483**	0.30818	**0.00127**	0.35732
Müller, ALD (GCL)	10000	**0.00468**	0.58947	0.00622	0.31122	**0.00545**	0.39470

Here Yamashiro and Müller do very well, with comparable bias to pure BF but somewhat greater RMS error. MCL, PIC, and the DCL family are unexceptional, mainly because of bias.

2.3.3. Model with parameter drift and trend

The results of the foregoing model with trend added are shown in the table on the following page. Here T represents the trend factor (by accident period, per annum) used to detrend the data entering the calculations and restore the trend to the results. The trend in the data is known: frequency trend of 1.01 multiplied by severity trend of 1.03, or approximately 1.04, which we take for T. Normally T must be estimated from exogenous data, such as industry studies or internal studies involving entire lines of business, so that it may not be as successful in explaining the trend in each loss history as is the case here.

The top half of this table shows selected results with no adjustment for trend, i.e. with T=1. Since there *is* trend in the data, *and this in the accident-year direction*, we would expect a major impact on the Cape Cod and pure PLR estimates (since the average loss ratios are derived mainly from the early years but applied mainly to the later years), some impact on the "PLR chain ladder" estimate (since PLR's at the later lags are based on earlier accident years only), and practically no effect on the chain ladder and related estimates (since development factors are independent of accident-year trends). The results confirm these expectations: in the presence of trend, the Cape Cod and pure PLR estimates become unreliable because of bias, while their associated chain-ladder estimates remain unreliable because of variance. Interestingly, BDCL and incurred MCL appear to be robust against accident-year trend.

A solution to the breakdown of these methods in the presence of trend is to detrend the data entering the estimators and restore trend to the output. This is shown in the bottom half of the table, with trend factor of 1.04 nearly matching the trend in the data. The trend adjustment reduces the bias of most of the traditional non-chain-ladder estimates, along with BDCL, the incurred MCL, and the two Affine models, to insignificant levels.

		Paid		Incurred		Combined	
Estimator	N	Bias	Error	Bias	Error	Bias	Error
No Adjustment for Trend (T=1)							
BF, G=1 (pure Cape Cod)	10000	-0.16440	0.41168	-0.10033	0.27943	-0.13237	0.33200
BF, G=0 (pure chain-ladder)	10000	0.22813	1.85283	**-0.00202**	0.37807	0.11306	1.02522
PLR BF, G=1 (pure PLR)	10000	-0.17779	0.41660	-0.10731	0.28426	-0.14255	0.33809
PLR BF, G=0 (PLR CL)	10000	-0.06262	0.71149	-0.04257	0.31413	-0.05259	0.46434
Quarg & Mack, MCL	10000	-0.07166	0.41672	**0.00432**	0.42136	-0.03367	0.38922
Merz & Wütrich, PIC	9995	0.10523	1.46415	0.05450	0.38060	0.06578	0.48301
Yamashiro, RC	9987	-0.02438	0.41462	-0.02438	0.41462	-0.02438	0.41462
Agbeko et al, DCL and IDCL	10000	0.21060	1.76039	-0.04401	0.40594	0.08330	0.98823
Agbeko et al, BDCL	10000	**-0.00085**	0.41540				
Müller, ALD (GLR)	10000	-0.15716	0.50634	-0.06729	0.30625	0.11222	0.35909
Müller, ALD (GCL)	10000	-0.15108	0.53538	-0.06440	0.30553	0.10774	0.36917
Adjusted for Trend, T=1.04							
BF, G=1 (pure Cape Cod)	10000	**-0.00082**	0.40079	-0.00639	0.27275	**-0.0036**	0.32223
BF, G=0 (pure chain-ladder)	10000	0.22824	1.86232	**-0.00362**	0.37674	0.11231	1.02909
PLR BF, G=1 (pure PLR)	10000	**-0.00293**	0.40293	-0.00633	0.27640	**-0.00463**	0.32629
PLR BF, G=0 (PLR CL)	10000	**-0.00698**	0.77459	-0.02431	0.32241	-0.01565	0.49625
Quarg & Mack, MCL	10000	-0.07355	0.41714	**0.00388**	0.42291	-0.03484	0.38979
Merz & Wütrich, PIC	9995	0.11007	1.46111	0.06256	0.38516	0.06722	0.48334
Yamashiro, RCy	9988	-0.02515	0.41279	-0.02448	0.41255	-0.02482	0.41265
Agbeko et al, DCL and IDCL	10000	0.21061	1.76820	-0.04551	0.40471	0.08255	0.99122
Agbeko et al, BDCL	10000	**-0.00228**	0.41454				
Müller, ALD (GLR)	10000	**0.00406**	0.48932	**-0.00444**	0.29559	**-0.00019**	0.34941
Müller, ALD (GCL)	10000	**0.00665**	0.51633	**-0.00457**	0.29526	**0.00104**	0.35832

2.3.4. Model with parameter drift, trend, and wider case-reserve errors

The results of the model with more pronounced case reserve errors (including bias at various stages between reporting and settlement) are shown in the table below.

		Paid		Incurred		Combined	
Estimator	N	Bias	Error	Bias	Error	Bias	Error
Adjusted for Trend, T=1.04							
BF, G=1 (pure Cape Cod)	10000	**0.00168**	0.40504	-0.00718	0.29232	**-0.00275**	0.32956
BF, G=0 (pure chain-ladder)	10000	0.23182	1.74620	**0.00044**	0.41318	0.11613	0.97570
PLR BF, G=1 (pure PLR)	10000	**-0.00054**	0.40706	-0.00737	0.29453	**-0.00396**	0.33323
PLR BF, G=0 (PLR CL)	10000	**0.00627**	0.78615	-0.02157	0.35148	**-0.00765**	0.50273
Quarg & Mack, MCL	10000	-0.05158	0.46547	0.01408	0.47317	-0.01875	0.44511
Merz & Wütrich, PIC	9996	0.10947	1.35860	0.08214	0.41933	0.08633	0.52803
Yamashiro, RC	9993	-0.01208	0.45257	-0.01156	0.45254	-0.01182	0.45255
Agbeko et al, DCL and IDCL	10000	0.21425	1.67091	-0.03810	0.43509	0.08807	0.94669
Agbeko et al, BDCL	10000	**0.00185**	0.43844				
Müller, ALD (GLR)	10000	**0.00276**	0.46467	**-0.00334**	0.33328	**-0.00029**	0.35894
Müller, ALD (GCL)	10000	**0.00599**	0.51558	**-0.00324**	0.33840	**0.00138**	0.37621

As expected, the incurred results display somewhat greater error than before. Note that the sampled loss histories are different from the foregoing example in both their paid and incurred triangles, rather than just the incurred, because simulating the bias in case reserves at various maturities requires additional samples from the underlying random number generator.

2.3.5. Complex model

The results of the complex model (multiple lines and trends, etc) are shown below. The RMS errors here are affected in opposite directions by the increased number of claims from the addition of a second line, and the fact that this "Line 2" has a much longer payment lag distribution than the "Line 1" inherited from the previous set of tests. Incurred MCL shows very little bias, as do both of the incurred ALD estimates, although most practitioners would probably prefer the Cape Cod or PLR estimates if only for their smaller RMS error.

The adjustment for trend is based on the total trends of Line 1 and Line 2, about 1.04 and 1.06, respectively. With a bit of experimentation we found that 1.055 produced satisfactory results; this is close to the trend of Line 2, probably because the long lags of that line increase its importance to the reserves. Of course, an actuary with a single loss history would not have the luxury of such experimentation!

		Paid		Incurred		Combined	
Estimator	N	Bias	Error	Bias	Error	Bias	Error
Adjusted for Trend, T=1.055							
BF, G=1 (pure Cape Cod)	10000	**0.00663**	0.54090	**-0.00027**	0.23647	**0.00318**	0.34844
BF, G=0 (pure chain-ladder)	10000	0.11505	1.25797	-0.00940	0.29739	0.05282	0.69530
PLR BF, G=1 (pure PLR)	10000	**0.00365**	0.50503	**0.00111**	0.24793	**0.00238**	0.34118
PLR BF, G=0 (PLR CL)	10000	-0.02536	0.66339	-0.02086	0.26719	-0.02311	0.40632
Quarg & Mack, MCL	10000	-0.02217	0.56480	**0.00093**	0.32481	-0.01062	0.39514
Merz & Wütrich, PIC	9999	0.06883	1.13850	0.05355	0.29560	0.04657	0.53196
Yamashiro, RC	9990	-0.04252	0.36911	-0.03598	0.35525	-0.03925	0.35873
Agbeko et al, DCL and IDCL	10000	0.09741	1.20636	-0.02855	0.30451	0.03443	0.67329
Agbeko et al, BDCL	10000	-0.04109	0.35145				
Müller, ALD (GLR)	10000	**0.01408**	1.00858	**0.00366**	0.28383	**0.00887**	0.56561
Müller, ALD (GCL)	10000	**0.01506**	0.98407	**0.00253**	0.27924	**0.00879**	0.55334

Effect of triangle size. To test how the size of each sample point (based on expected frequency) affects the performance of some of these estimators, we combined each five successive original sample points into a new sample point five times larger, and similarly with ten original sample points. The results are shown on the next page.

For this complex simulation, the immaturity of the triangles fed to the estimators is slightly more material than with the earlier sets of loss histories. The ratio of losses paid after lag 10 years to the total paid runoff is about 0.6%, and the ratio of losses recognized in the incurred triangle after lag 10 years to the total paid runoff is about 0.03%. These losses, between lags 10 and 13, are included in the runoff, but the known portion is not supplied to the estimators, which all work with 10 x 10 triangles. This contributes a small negative bias to all the results in this table.

The incurred PIC method and the ALD methods, both paid and incurred, do remarkably well with respect to bias and RMS error with these larger triangles, but are still not as close to zero, nor as tightly distributed, as either the Cape Cod or the pure PLR.

		Paid		Incurred		Combined	
Estimator	N	Bias	Error	Bias	Error	Bias	Error
Frequency x 5; Adjusted for Trend, T=1.055							
BF, G=1 (pure Cape Cod)	2000	**0.00114**	0.23277	**-0.00076**	0.11054	**0.00019**	0.15410
BF, G=0 (pure chain-ladder)	2000	**-0.00582**	0.45188	-0.01961	0.13727	-0.01272	0.26067
PLR BF, G=1 (pure PLR)	2000	**0.00365**	0.22377	**0.00111**	0.11335	**0.00238**	0.15262
PLR BF, G=0 (PLR CL)	2000	-0.02891	0.34035	-0.01920	0.12990	-0.02406	0.20689
Quarg & Mack, MCL	2000	-0.02061	0.20077	-0.01456	0.14819	-0.01758	0.16301
Merz & Wütrich, PIC	2000	**-0.00193**	0.45540	**0.00268**	0.13141	**0.00084**	0.19496
Yamashiro, RC	2000	-0.03166	0.17463	-0.02867	0.15947	-0.03016	0.16461
Agbeko et al, DCL and IDCL	2000	**-0.01675**	0.43667	-0.01972	0.13708	-0.01823	0.25302
Agbeko et al, BDCL	2000	-0.04393	0.17277				
Müller, ALD (GLR)	2000	**0.00673**	0.29526	**0.00237**	0.13719	**0.00455**	0.18629
Müller, ALD (GCL)	2000	**0.00798**	0.29636	**0.00170**	0.13658	**0.00484**	0.18623
Frequency x 10; Adjusted for Trend, T=1.055							
BF, G=1 (pure Cape Cod)	1000	**-0.00032**	0.16097	**-0.00088**	0.07791	**-0.00060**	0.10687
BF, G=0 (pure chain-ladder)	1000	-0.02299	0.31188	-0.02117	0.09742	-0.02208	0.18114
PLR BF, G=1 (pure PLR)	1000	**0.00365**	0.15554	**0.00111**	0.07942	**0.00238**	0.10585
PLR BF, G=0 (PLR CL)	1000	-0.02930	0.25414	-0.01902	0.0928	-0.02416	0.15315
Quarg & Mack, MCL	1000	-0.02374	0.13615	-0.01854	0.10174	-0.02114	0.11094
Merz & Wütrich, PIC	1000	**-0.01862**	0.31366	-0.00712	0.09212	-0.01040	0.13684
Yamashiro, RC	1000	-0.03192	0.12804	-0.02890	0.11519	-0.03041	0.11991
Agbeko et al, DCL and IDCL	1000	-0.03258	0.30158	-0.02117	0.09742	-0.02688	0.1761
Agbeko et al, BDCL	1000	-0.04417	0.12852				
Müller, ALD (GLR)	1000	**0.00444**	0.19669	**0.00356**	0.09785	**0.00400**	0.12532
Müller, ALD (GCL)	1000	**0.00488**	0.19704	**0.00321**	0.09677	**0.00404**	0.12514

Prediction of latest year only. The above tables compare the reserve by each estimator with the simulated runoff for all accident years combined. For some purposes, particularly ratemaking, we prefer comparisons by accident year. The ultimate losses by year are usually of greater interest than the reserves. In the earliest years, even large relative errors in forecasting the reserves may have negligible impact on the estimated ultimate losses, but the reserves dominate the more relevant recent years, so we continue to compare estimators by the reserves. For the latest accident year:

		Paid		Incurred		Combined	
Estimator (Year 10 only)	N	Bias	Error	Bias	Error	Bias	Error
Adjusted for Trend, T=1.055							
BF, G=1 (pure Cape Cod)	10000	0.01902	0.49923	**0.00598**	0.43331	0.01250	0.45524
BF, G=0 (pure chain-ladder)	10000	0.30381	3.00878	-0.01468	0.67406	0.14456	1.65219
PLR BF, G=1 (pure PLR)	10000	0.01813	0.49517	**0.00736**	0.43604	0.01275	0.45673
PLR BF, G=0 (PLR CL)	10000	**-0.01365**	1.30555	-0.03976	0.55982	-0.02670	0.80583
Quarg & Mack, MCL	10000	-0.04614	0.79649	**0.00755**	0.75418	-0.01930	0.74678
Merz & Wütrich, PIC	9999	0.13852	2.56507	0.11288	0.66915	0.12618	0.86074
Yamashiro, RC	9990	-0.03326	0.79578	-0.03162	0.79614	-0.03244	0.79585
Agbeko et al, DCL and IDCL	10000	0.27420	2.93756	-0.06386	0.70457	0.10517	1.62928
Agbeko et al, BDCL	10000	-0.03606	0.66818				
Müller, ALD (GLR)	10000	0.01931	0.64308	**0.00719**	0.48433	0.01325	0.51004
Müller, ALD (GCL)	10000	0.02017	0.74935	**0.00678**	0.48230	0.01347	0.54499

As expected, the RMS errors are much greater for a single year than for the total reserve. They are enough smaller for Cape Cod and pure PLR to give little reason to prefer any of the other estimators for this purpose.

Gluck factors between 0 and 1. It was mentioned above that Gluck factors other than 1 (Cape Cod) or 0 (chain ladder) are primarily useful in an environment of changing loss ratios. But even with modest changes, such as the parameter drift that remains after we adjust for the trend in our complex model, the choice of Gluck factor may be important when projecting the latest year.

		Paid		Incurred		Combined	
Estimator (Year 10 only)	N	Bias	Error	Bias	Error	Bias	Error
Adjusted for Trend, T=1.055							
BF, G=1 (pure Cape Cod)	10000	0.01902	0.49923	**0.00598**	0.43331	0.01250	0.45524
BF; G=0.8	10000	0.01206	0.50566	**-0.00171**	0.43134	**0.00518**	0.45158
BF; G=0.6	10000	**0.00558**	0.53518	-0.00955	0.44073	**-0.00198**	0.46099
BF; G=0.4	10000	**0.00497**	0.60349	-0.01520	0.46224	**-0.00511**	0.48933
BF; G=0.2	10000	0.02621	0.84881	-0.01786	0.51221	**0.00418**	0.59946
BF, G=0 (pure chain-ladder)	10000	0.30381	3.00878	-0.01468	0.67406	0.14456	1.65219
PLR BF, G=1 (pure PLR)	10000	0.01813	0.49517	**0.00736**	0.43604	0.01275	0.45673
PLR BF; G=0.8	10000	**0.00888**	0.49202	**-0.00110**	0.43198	**0.00389**	0.44922
PLR BF; G=0.6	10000	**-0.00260**	0.50286	-0.01063	0.43653	**-0.00662**	0.45058
PLR BF; G=0.4	10000	-0.01286	0.53788	-0.01930	0.44778	-0.01608	0.46283
PLR BF; G=0.2	10000	-0.01851	0.64544	-0.02778	0.47258	-0.02314	0.50572
PLR BF, G=0 (PLR CL)	10000	**-0.01365**	1.30555	-0.03976	0.55982	-0.02670	0.80583

The RMS errors generally increase as the Gluck factor decreases, and the estimator takes on more of the characteristics of chain-ladder. But the minimum absolute bias is with G=0.8 for the incurred estimators and with G=0.4 or 0.6 for the paid; under the right conditions this simple means of capturing changes over time in the ELR can outperform BF Cape Cod.

2.3.6. Correlations and combinations of estimators

It is common practice for actuaries to use a linear combination of two or more dissimilar estimates, with coefficients selected by judgment, in the expectation that the combination will give a more stable estimate than any one of its components. Our "combined" columns in the above tables demonstrate the commonly used average of paid and incurred, for most of the estimators studied.

To explore combinations other than just paid and incurred versions of the same estimator, we measured the correlations between the reserves, and between the prediction errors, of all pairs of estimates. As might be expected, the reserves were highly positively correlated, as they are driven directly by the losses emerged to date, which are common to all estimators. The prediction errors between similar estimators (e.g. BF and PLR) were also highly positively correlated; almost all the other pairs of prediction errors were positively correlated, giving little opportunity for reducing RMSE drastically via a linear combination.

As a modest example where a combination of estimators may reduce the RMS error, using the simple data set, the paid Cape Cod estimate has bias of 0.00036 and RMS error of 0.39784; the incurred MCL model has bias of 0.02461 and RMS error of 0.51150; the correlation of the errors in these two estimators is about 21%. Taking a straight average of the two gives bias of 0.01248 and RMS error of 0.35402, smaller than either of the components. For another example, combining the paid Cape Cod estimate with the paid Yamashiro Recursive Credibility estimate (bias -0.00742, RMSE 0.40880) reduces the final RMSE to 0.33625, again smaller than either component. But better and simpler estimates are at hand, such as the incurred Cape Cod.

Part of the variability of the estimators studied here is attributable to egregious outliers that would be easily detected and rejected in practical loss reserving, but that lose themselves among the 10,000 sample points of our simulations. The problem may not be easily spotted in the data: for example, it is unremarkable for large cumulative losses at the late lags to move very slightly, but this sometimes produces ill-conditioned matrices to be inverted in the ALD methods. Minimizing the influence of outliers suggests running several different estimators and taking the median, or, more generally, an average of the estimators remaining after discarding the one or two greatest, and one or two least, results for each sample point. For example, we might consider combinations of the recently published estimators applied to the complex data set:

Estimator	N	Bias	Error
Incurred BF, G=1 (pure Cape Cod)	10000	**-0.00027**	0.23647
Median of incurred MCL, PIC, RC, DCL, and ALD GLR	10000	0.00003	0.28716
Average of central 4 of incurred MCL, PIC, RC, DCL, and both ALD's,	10000	0.00122	0.26646
Average of central 2 of incurred MCL, PIC, RC, DCL, and ALD's	10000	0.00183	0.27577
Median of BDCL and paid and incurred MCL, PIC, RC, DCL, and ALD's	10000	-0.02317	0.38261
Avg of central 11 of BDCL and paid and inc'd MCL, PIC, RC, DCL, and ALD's	9989	-0.00775	0.35236
Avg of central 9 of BDCL and paid and inc'd MCL, PIC, RC, DCL, and ALD's	10000	-0.01686	0.37178

Such combinations work, but they do not work miracles. The incurred combinations remain superior to the combinations including paid estimators, and even the best of these combinations, while better than many of their components, are not superior to the incurred Cape Cod.

3. CONCLUSIONS

Our main conclusions are that simulation remains a valuable means of evaluating loss reserve estimators, that the CAS Loss Development Simulator (even in its more limited prototype version) is a useful tool for conducting simulations, that applying this tool to traditional and proposed loss reserving methods can help test their practical usefulness, and that, in such tests, the early, parsimonious, methods of Bornhuetter-Ferguson, Stanard, Bühlmann, and Gluck generally hold their own very well.

Some secondary conclusions relate to the particular methods tested here:

Pure chain-ladder applied to paid losses is unsatisfactory for its extreme RMS error and apparent positive bias. This appears to be exacerbated here by the small number of claims in each triangle; the performance improves noticeably, by comparison with other estimators, with larger triangles.

Pure chain-ladder applied to incurred losses is much better than for paid losses, but is still generally outperformed by pure Cape Cod.

The Partial Loss Ratios estimators generally show very little bias. The pure PLR is comparable in bias and RMS error to the pure Cape Cod, while the "PLR chain-ladder" easily outperforms the conventional chain-ladder.

For the particular data sets simulated, most of the incurred estimates performed much better than the corresponding paid estimates, when measured by RMS error, and often when measured by absolute bias. This may reflect the modest case reserve error assumed here.

The mean of the paid and incurred estimates has no obvious advantage over the incurred estimate alone. In most cases the errors from the paid and incurred estimates using the same estimator are highly correlated.

Quarg and Mack's Munich Chain Ladder, Merz and Wüthrich's Paid-Incurred Chain method, Yamashiro's Recursive Credibility model, Agbeko's DCL, IDCL, and BDCL models, and Müller's Affine Loss Development models all performed adequately – in some cases very well - but in general were less successful than their simpler conventional counterparts. All of them require adjustment, and/or the omission of some sample points, to perform properly on small data sets.

The recently published estimators tested here are elegant and creative approaches to two of the most important open issues in loss reserving: how to combine paid and incurred information, and how to obtain reserves as a posterior distribution rather than a deterministic estimate. But all depend on distributional models of the loss process in aggregate, rather than its detailed components. When the detailed components are themselves modeled and combined via simulation, the resulting aggregate loss process is unlikely to satisfy these distributional models. Real-world processes, driven by even more detailed components than our simulations, are likely to depart even further from the distributional assumptions underlying methods such as PIC and MCL.

In summary, actuaries are well justified in continuing to prefer estimators such as Bornhuetter-Ferguson with Cape Cod ELR's, or ELR's using Gluck decay factors, to either the chain-ladder family or to more complex published models derived from chain-ladder. The simplicity of these estimators appears to be their strength.

4. REFERENCES

[1] Agbeko, Tony, Munir Hiabu, Maria Dolores Martinez-Miranda, Jens Perch Nielsen, and Richard Verrall, Validating the Double Chain Ladder Stochastic Claims Reserving Method, *Variance* Volume 8 Issue 2, 2014: 138-160

[2] Bornhuetter, Ronald and Ronald Ferguson, The Actuary and IBNR, *PCAS* LIX, 1972: 181-195

[3] Bühlmann, Hans, Vereinigung Schweizerischer Versicherungsmathematiker / Association des Actuaires Suisses, Ecole d'été 1983, Estimation of IBNR Reserves by the Methods Chain Ladder, Cape Cod, and Complementary Loss Ratio, unpublished

[4] CAS Loss Simulation Model Working Party, Modeling Loss Emergence and Settlement Processes, *CAS E-Forum*, 2011 Winter Volume 1: 1-124

[5] Gluck, Spencer M., Balancing Development and Trend in Loss Reserve Analysis, *PCAS* LXXXIV, 1997: 482-532

[6] Merz, Michael and Mario V. Wüthrich, Paid-incurred chain claims reserving method, *Insurance: Mathematics and Economics*, 46, 2010: 568-579

[7] Müller, Thomas, Projection for Claims Triangles by Affine Age-to-Age Development, *Variance* Volume 10, Issue 1: 121-144

[8] Quarg, Gerhard and Thomas Mack, Munich Chain Ladder: A Reserving Method that Reduces the Gap between IBNR Projections Based on Paid Losses and IBNR Projections Based on Incurred Losses, *Blätter der Deutschen Gesellschaft für Versicherungs- und Finanzmathematik*, volume 26, number 4, 2004: 597-630. Reprinted in *Variance*, Volume 2/Issue 2, 2008: 266-299

[9] Shang, Kailan, Loss Simulation Model Testing and Enhancement, *CAS E-Forum*, 2011 Summer: 1-75

[10] Stanard, James N., A Simulation Test of Prediction Errors for Loss Reserve Estimation Techniques, *PCAS* LXXII: 124-153

[11] Vaughan, Richard L., *Some Extensions of J.N. Stanard's Simulation Model for Loss Reserving*, CAS Forum, Fall 1998

[12] Yamashiro, Marcus M., Recursive Credibility: Using Credibility to Blend Reserve Assumptions, *Variance* Volume 8 Issue 2, 2014: 105-137