Text mining and ruin theory: A case study of research on risk models with dependence *

Authors: Renata G. Alcoforado
– ISEG & CEMAPRE, Universidade de Lisboa;
Department of Accounting and Actuarial Sciences, Universidade Federal dePernambuco, Portugal & Brazil (alcoforado.renata@ufpe.br)

Alfredo D. Egídio dos Reis
– ISEG & CEMAPRE, Universidade de Lisboa
Portugal (alfredo@iseg.ulisboa.pt)

Abstract:

- This paper aims to analyze unstructured data using a text mining approach. The work was motivated in order to organize and structure research in Risk Theory. In our study, the subject to be analyzed is composed by 27 published papers of the risk and ruin theory topic, area of actuarial science. They were coded into 32 categories. For the purpose, all data was analyzed and figures were produced using the software *NVivo 11 plus*. Software *NVivo* is a specialized tool in analyzing unstructured data, although it is commonly used just for qualitative research. We used it for Quali-Quant analysis.

Key-Words:

- *Big data; Unstructured data; Text mining; Risk theory; Ruin theory; Dependence modeling.*

AMS Subject Classification:

- 49A05, 78B26.

*The opinions expressed in this text are those of the authors and do not necessarily reflect the views of any organization.

1. INTRODUCTION

As widely known, Big Data is an area of great development in statistics. We can define Big Data as "a phenomenon defined by the rapid acceleration in the expanding volume of high velocity, complex, and diverse types of data. Big Data is often defined along three dimensions – volume, velocity, and variety" (TechAmerica Foundation's Federal Big Data Commission, 2012).

According to Han *et al.* (2012) data mining is the process of mining through large amount of data to extract meaningful information, knowledge. It's also treated by many people as a synonym for knowledge discovery from data, simply KDD.

Text mining in an analogous manner as data mining, aims to extract information from data, but in this case the data comprehend to texts and do it through identification and exploration of interesting patterns (Feldman and Sanger, 2006). Accordingly to Aggarwal and Zhai (2012), the primary goal of text mining is analyzing information to discover patterns, going beyond information access to further help users analyze and digest information and facilitate decision making.

Text mining has been used as a form to extract knowledge from text, it has been applied to social media (see Corley *et al.* (2010), Zeng *et al.* (2010), Maynard *et al.* (2012), He *et al.* (2013), Mostafa (2013)), health science (see Chen *et al.* (2005), Cohen and Hersh (2005), Collier *et al.* (2008), Zweigenbaum *et al.* (2007), Hirschman *et al.* (2012)), in social sciences (see Peng *et al.* (2012)) and other fields.

Francis and Flynn (2010) show that text mining can be used to generate new information from the unstructured text data. Text mining can also be used to extract quantitative information, as Kim and Jun (2015) did to obtain a Gaussian copula regression model.

This paper was motivated to organize and structure our research in Risk Theory, the goal is to study this thematic in the most embracing, as well as profoundly, way. First, we need to know what has been studied in this topic so we selected the papers in the area and we aimed to extract knowledge from this database. We uploaded it in the software so it can be read for us.

The software can recognize patterns and present pertinent connections that otherwise we would miss and also spot the most pertinent papers in the area. The *NVivo* is usually used to qualitative analysis, but as Kim and Jun (2015) did in their paper, we also did a quali-quant analysis that evidence the ability to use this software for quantitative analysis and we expect that others researchers will do the same.

This paper is organized as follows: In Section 2 we speak about the collected data under analysis. Section 3 is about the coding of the data, the coding matrix, the relationship between the nodes, that is, we plotted the nodes hierarchically. Then, we present the cluster analysis for the nodes and the sources (papers), the comparison diagrams and to finalize, a structural matrix. To conclude, in Section 4 we write some final remarks.

2. THE DATA

Our data is composed by published scientific papers. For this particular study we chose a limited set, enough for our immediate purpose. Working with reference material in many aspects is no different from working with any other form of text. As it is in the form of research literature, it will contain author defined sections that can be compared across the references. Also, keywords are available. Therefore we can consider that this type of data is also likely to be more structured than information from an interview [see for example Bazeley and Jackson (2013)].

We chose a set of 27 scientific papers to be analyzed. We uploaded these 27 papers in the platform, coded and then analyzed all data. These papers are references for a particular research project in development in risk theory. These papers are: Afonso *et al.* (2017), Ammeter (1948); Asmussen and Albrecher (2010); Bergel and Egídio dos Reis (2016); Constantinescu *et al.* (2011); Constantinescu *et al.* (2012); Constantinescu *et al.* (2016); Czado *et al.* (2011); Frees and Wang (2006); Frees *et al.* (2011); Frees *et al.* (2016); Garrido *et al.* (2016); Gschlößl and Czado (2007); Jasiulewicz (2001); Jørgensen and Paes De Souza (1994); Krämer *et al.* (2013); Kreer *et al.* (2015); Li *et al.* (2015); Maume-Deschamps *et al.* (2017); Ni *et al.* (2014a); Ni *et al.* (2014b); Quijano Xacur and Garrido (2015); Renshaw (1994); Rolski *et al.* (1999); Schulz (2013); Shi *et al.* (2015) and Song *et al.* (2009).

Using the software, the first task we took was to build a *word cloud* composed by the the most pertinents words in our entire data base to use in our study. After removing all the verbs, articles and non-meaningful wording, the words are then gathered according to their *stem*, then search the frequency of words, making possible to obtain the *cloud* as shown in Figure 1. It's important to point out that the *word cloud* shows the essence of the data base, where the size matters.

In the coding we will present the figures in the order in which we elaborated them. First, as prior mentioned is the word cloud in Figure 1, which will contribute on the creation of the categories. Then, in Figure 2 is presented the Word Tree for the node "Aggregate claims model", that we obtain when coding the database.

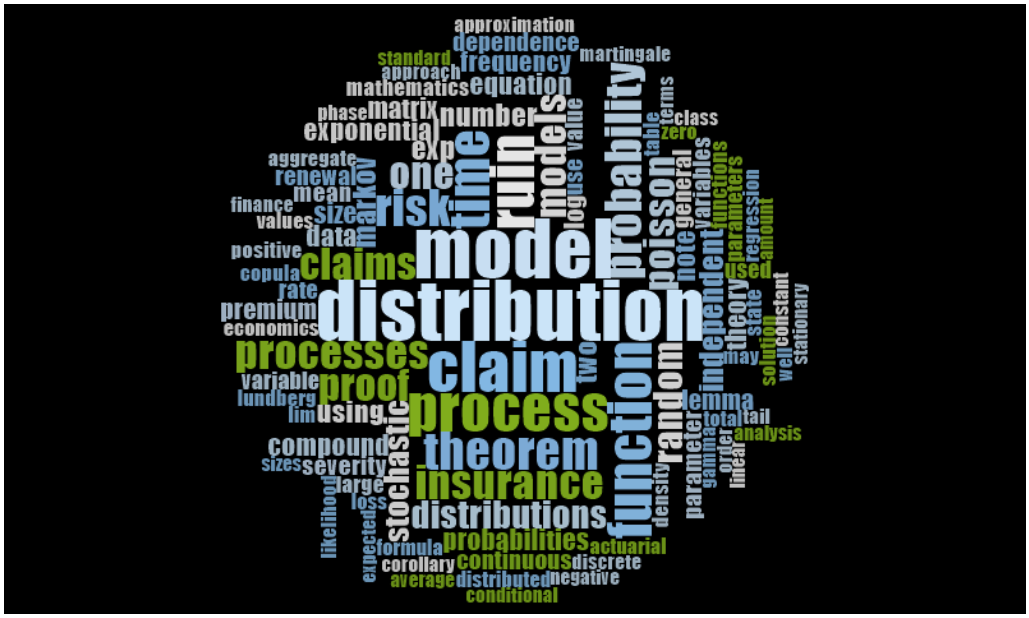

Figure 1: Word Cloud

In Figure 3 is a chart node coding for "Claim Severity", that derives from this specific category after coding the database. In sequence, we desire to see how each one of the categories fit hierarchically in the entire group of categories and also how they connect with one another, therefore we present them in Figure 4 and in Figure 5, respectively.

Then, we analyze first the categories and then the sources using cluster analysis, for the Cluster analysis of the categories we exhibit two figures, in Figure 6 is the circle graph and in Figure 7 is the dendrogram. As a result of the cluster analysis for the sources we display one dendrogram in Figure 8.

Posteriorly, we conclude from the cluster analysis and from the coding matrix the categories that are interesting to compare, hence we present in Figure 9 two comparison diagrams. Finally, we present a summarized framework matrix.

3. THE CODING AND ANALYSIS

A code is an abstract representation of a case. Corbin and Strauss (2008) says that we can think of coding as "mining" the data, digging beneath the surface to discover the hidden treasures contained within data. Accordingly to Bazeley and Jackson (2013), coding is one of several methods of working with and building knowledge about data.

Data mining assumes that the data is already stored in an structured way,

whereas text mining assumes that the data is unstructured and still needs coding, accordingly to Feldman and Sanger (2006).

In sequence, all the unstructured data was coded, building categories, that is, we put in each one of the categories the respective parts from text to be able to analyze it in a mathematical way. In other words, after coding we get a structure to be able to analyze with clusters and matrices. With that, we can plot the data now, this was not possible before. The categories were selected after extensive reading and observing the *word cloud*.

In our particular analysis the *codes* are: Actuarial; Aggregate Claims Model; Claim Frequency; Claim Severity; Compound Poisson; Conditional; Copulas; Covariates; Dependence; Exponential; Formula; Function; Gamma; Independence; Insurance; Joint Distribution; Loss; Markov; Martingale; Mixed-Poisson; Parameters; Prediction; Premium; Randomness; Regression; Renewal; Risk Theory; Ruin Probability; Simulation; Spatial; Stationary and Stochastic Process.

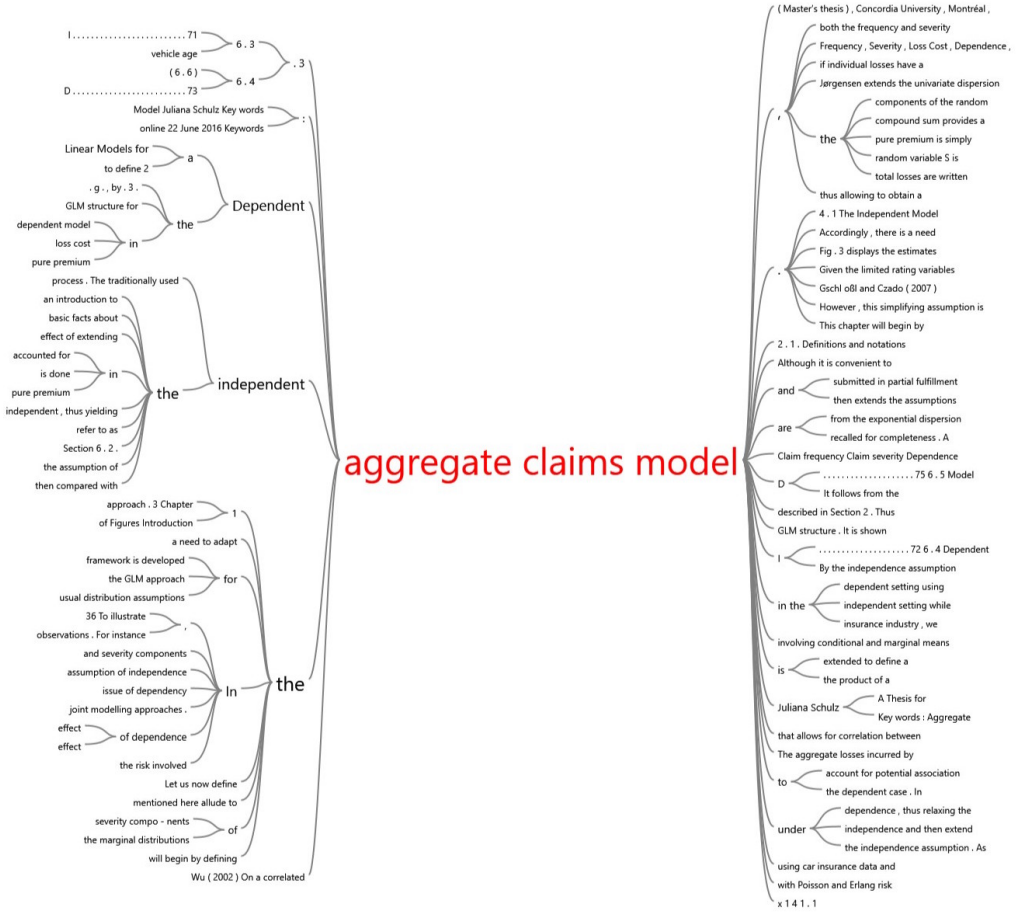

Figure 2: Word Tree - Aggregate Claims Model

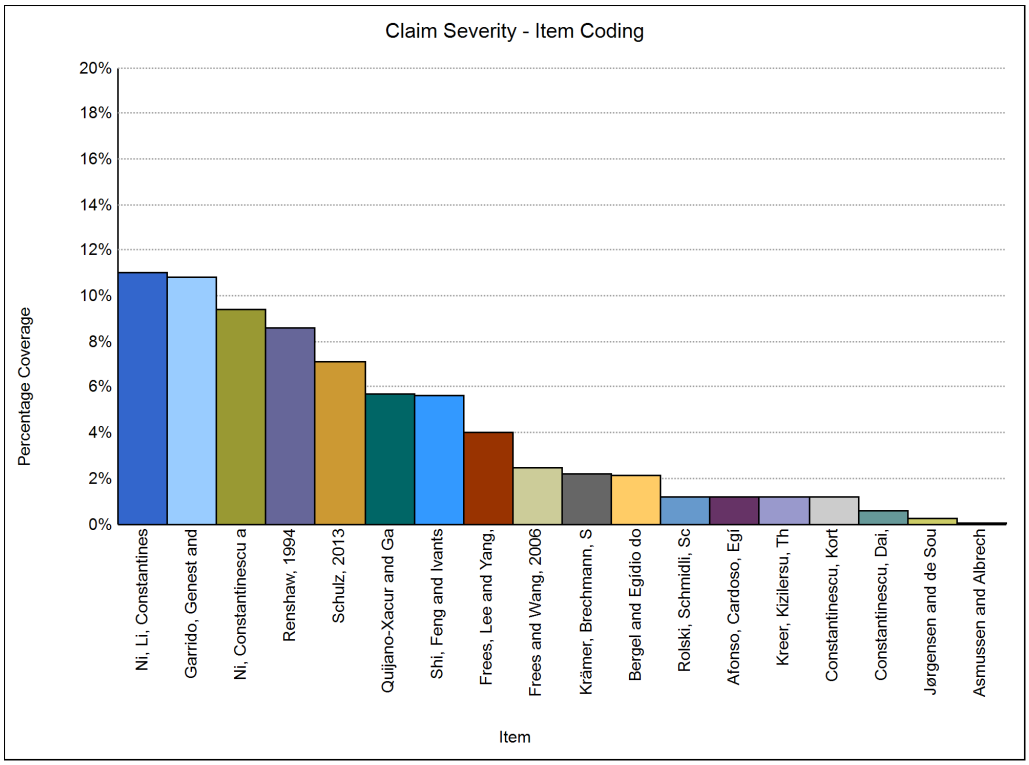

Figure 3: Chart Node Coding - Claim Severity

After the code and data organization, for each category is plotted a *word tree* to see the connection from that word (or expression) in the sentence where it belongs. An example is given in Figure 2, we can observe how the "aggregate claims model" fits in the sentence. In this case, authors are talking mostly about the "dependent" and the "independent" aggregate claims model. They also talk about the "issue of dependence", "assumption of independence", "the marginal distributions", "the structure" and "the effect of extending" the aggregate claims model.

For every category is plotted a chart node coding that presents the sources from our database that address the most and the importance that each paper from the database gives to that code. In Figure 3 we can observe which authors and in which papers the category "Claim Severity" is included. So, we can distinguish the author Constantinescu from our database since four of the papers that address the most to "Claim Severity" are written by her, including the first one.

We plotted the *nodes*, or categories, hierarchically presented in Figure 4 to observe which categories are most frequent, the most important among the data available.

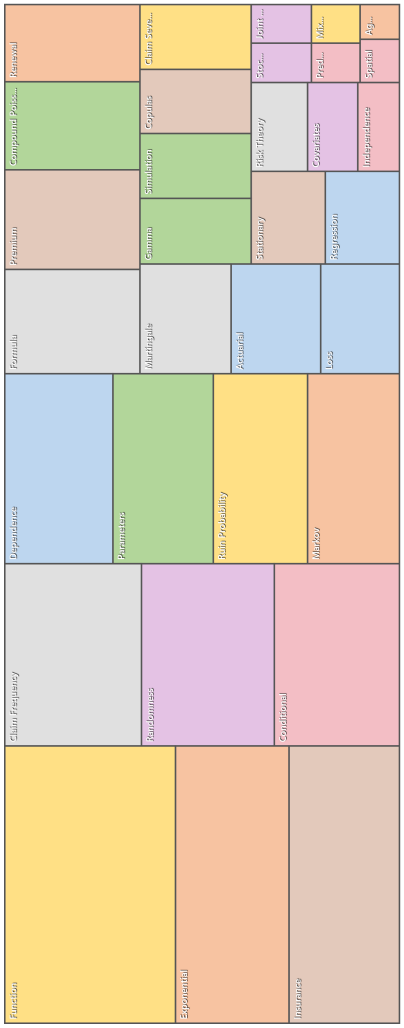

Figure 4: Nodes Hierarchically

In Figure 4 we can observe how the category that have the most importance is "Function", then followed by "Exponential" and then "Insurance". Another fact to point out is how "Claim frequency" is more important hierarchically than "Claim Severity", which is in line with the fact the most motor insurance models don't consider the Claim severity.

The authors when trying to capture the dependence between claim frequency and severity can use a "Regression" approach in which use one variable as a "Covariate" in the others regression or they can use a "Copula" approach. In the Figure 4 we can see how although they are almost the same size, "Regression" is still a bigger category.

Also, they can use a distribution to model when trying to capture the dependence, that distribution can be in hierarchically order: "Exponential", "Compound Poisson", "Gamma" and "Mixed Poisson". We can observe that stochastic processes are also very used. So, we can point out the following categories that fits into that description: "Markov", "Martingale", "Stationary" and "Stochastic Process" itself. As our database consists in authors that are trying to capture dependence between the two variables in some way, it's also important to mention how the code "Dependence" is more relevant then "Independence".

Our target as a research topic is to be able to calculate "Premium" and "Ruin Probability", although both categories have almost the same size it's important to mention that in one hand 22 out of the 27 papers address "Premium", while 10 papers address to "Ruin Probability". On the other hand, from those 10, there are 614 coded references for "Ruin Probability" and in those 22 papers there are 464 coded references for "Premium". To conclude the analysis of Figure 4, the method used to calculate these two quantities "Ruin Probability" and "Premium" as mentioned above can also be theoretical through "Formulas" or numerical through "Simulation". The former is the one that is the most sought in this database.

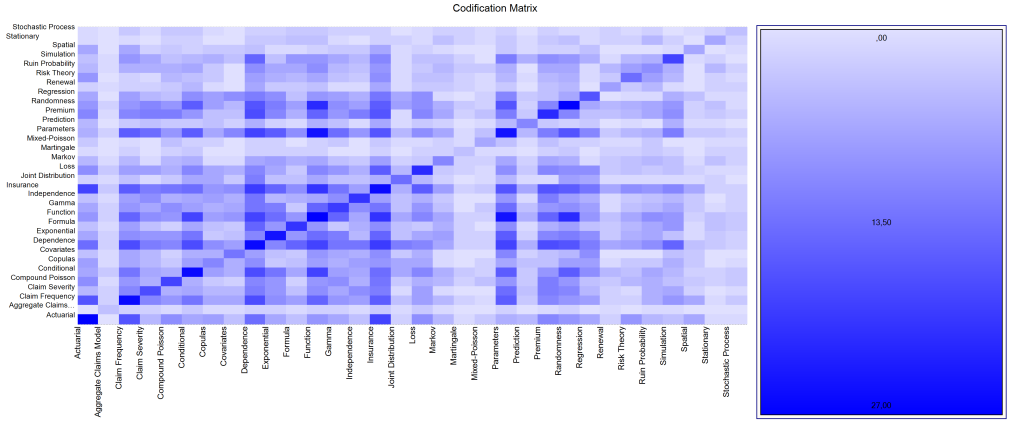

Figure 5: Coding Matrix - Heat

After, we constructed the *coding matrix* presented on the Figure 5, which shows both in numbers and in graph what is the relationship between the coded categories. In this matrix the colors are meaningful, the darker the color, the more codes are represented in the other coded references. In percentage, until 1% is white, from 1% to 10% is light blue, from 10% to 20% is a shade darker as we can observe between "Claim Frequency" and "Spatial", from 20% to 30% is another shade darker as we can observe between the node "Premium" in the row and column. The darkest blue means that it's between 30% and 40% as in "Exponential" in the row and column.

Although we may think to be symmetric, this matrix coding is not symmetric. It would be if we used the numbers, but the numbers are not as important as the percentage of the total for that category. Each cell content is the column percentage of coded references, and it's not symmetric because of the way the data was collected. That is, the papers are about dependence between the claim frequency and severity random variables, as a consequence the codes are going to reference more dependence than the other way around.

So, for instance when we consider "Copulas" and "Dependence", dependence is in 10.88% of the coded references from Copulas, and Copulas are in only 4.68% coded references of the Dependence category. Another case is "Premium" and "Insurance", "Insurance" are represented in 10.41% of the "Premium" coded references while "Premium" are represented in 6.93% of the "Insurance" category.

The *cluster analysis* was afterwards performed in cluster by word similarity using Pearson's correlation coefficient as the similarity measure. We made it for both the categories and sources to see how they relate. The cluster analysis for the nodes is presented in Figures 6 and 7, a circle graph in Figure 6 and a dendrogram in Figure 7.

In the circle graph in Figure 6 the colors represent the clusters and the lines represent the connection between the nodes, the more and the thicker are the lines, the higher is Pearson's correlation coefficient. We can observe an asymmetry to the right that means that the nodes on the right have a higher correlation.

Referring now to Figure 7, in this dendrogram we can observe 10 clusters for the 32 nodes represented by the colors and the branches. The following categories before mentioned for stochastic processes are in one cluster together with "Claim Frequency", since the claim frequency is usually considered as an stochastic process. The coefficient between "Stochastic Process" and "Martingale" is 0.815.

The cluster with the highest similarity is the one that comprehends "Function", "Conditional", "Exponential", "Formula", "Randomness" and "Renewal", the coefficient between "Function" and "Conditional" is 0.848, between "Formula" and "Conditional" is 0.820, between "Exponential" and "Conditional" is 0.817.

The categories "Independence" and "Dependence" present a 0.806 coeffi-

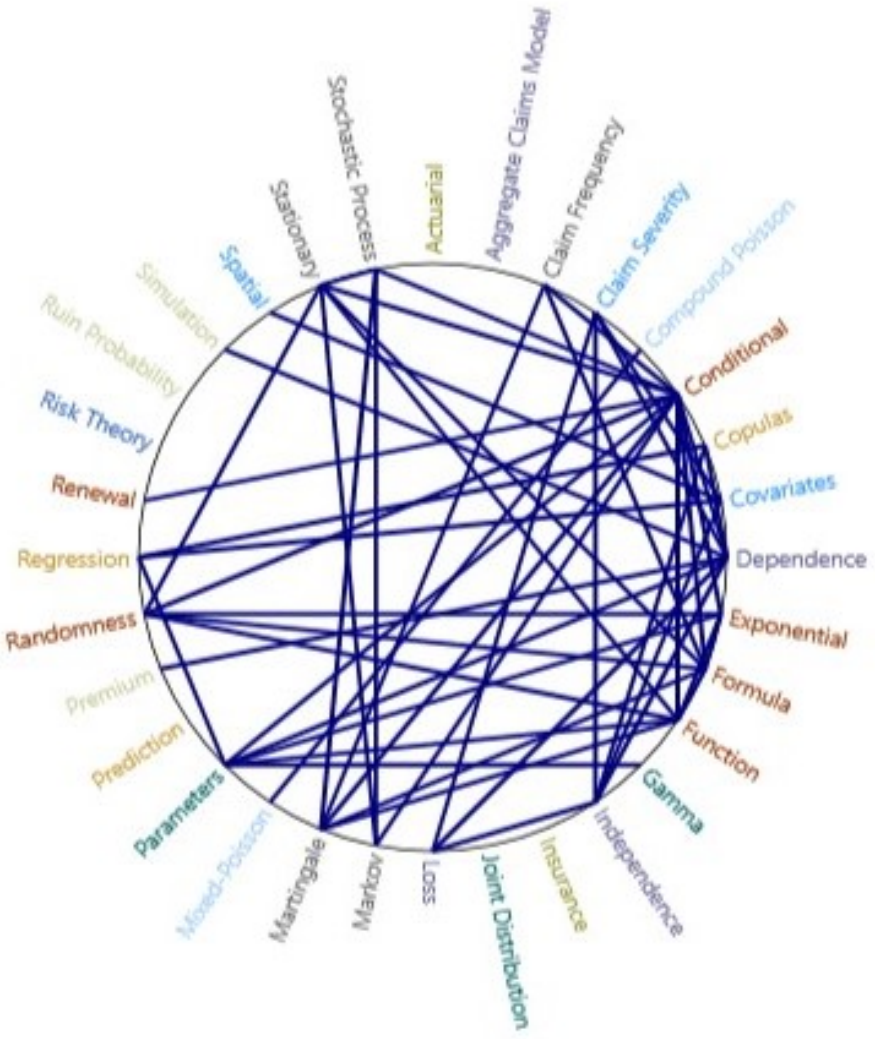

Figure 6: Cluster Analysis of the Nodes - Circle Graph

cient and are clustered together. "Ruin Probability" and "Premium" present a 0.645, both are clustered with "Simulation". "Claim Severity" and "Claim Frequency" a 0.521, "Claim Severity" is in a cluster with "Covariates" and "Spatial" while "Claim Frequency" is in the first cluster mentioned. "Simulation" and "Formula" a 0.616 and are in different clusters. And finally, "Copulas" and "Regression" a 0.793 and both are in the same cluster (the yellow).

Cluster analysis for the sources from out database was also plotted and is presented in Figure 8. From Figure 8 we can observe the clusters accordingly to colors and branches. There are three big clusters that comprehend 18 papers. The clusters are build using the complete linkage hierarchical clustering algorithm, also known as farthest neighbor clustering.

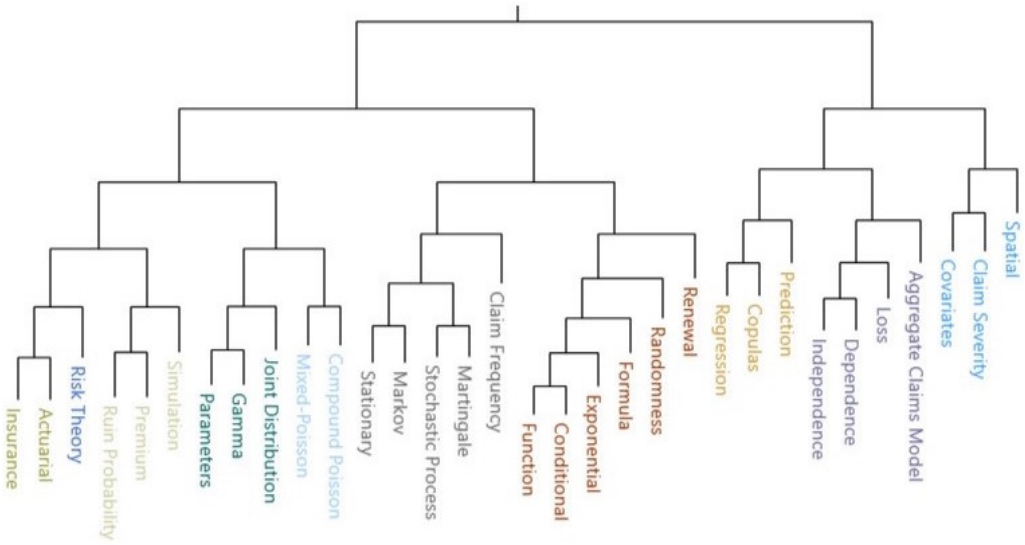

Figure 7: Cluster Analysis of the Nodes - Dendrogram

The higher correlation coefficients are in the middle cluster, the green one, between Frees *et al.* (2016) and Shi *et al.* (2015) is 0.91, between Shi *et al.* (2015) and Czado *et al.* (2011) is 0.88 and if we consider Frees *et al.* (2011), Frees *et al.* (2016), Shi *et al.* (2015), Afonso *et al.* (2017) and Czado *et al.* (2011) the correlation coefficient between two of them at a time goes from 0.83 to 0.91.

The blue cluster groups the papers Maume-Deschamps *et al.* (2017), Asmussen and Albrecher (2010), Rolski *et al.* (1999), Constantinescu *et al.* (2011), Bergel and Egídio dos Reis (2016) and Li *et al.* (2015), the coefficients between those vary from 0.65 (the farthest sources, Maume-Deschamps *et al.* (2017) and Li *et al.* (2015)) to 0.86, coefficient between Rolski *et al.* (1999) and Asmussen and Albrecher (2010).

We also plotted comparison diagrams. We present in Figure 9 the diagram comparing "Copulas and Covariates" on the left and the diagram comparing "Formula and Simulation" on the right.

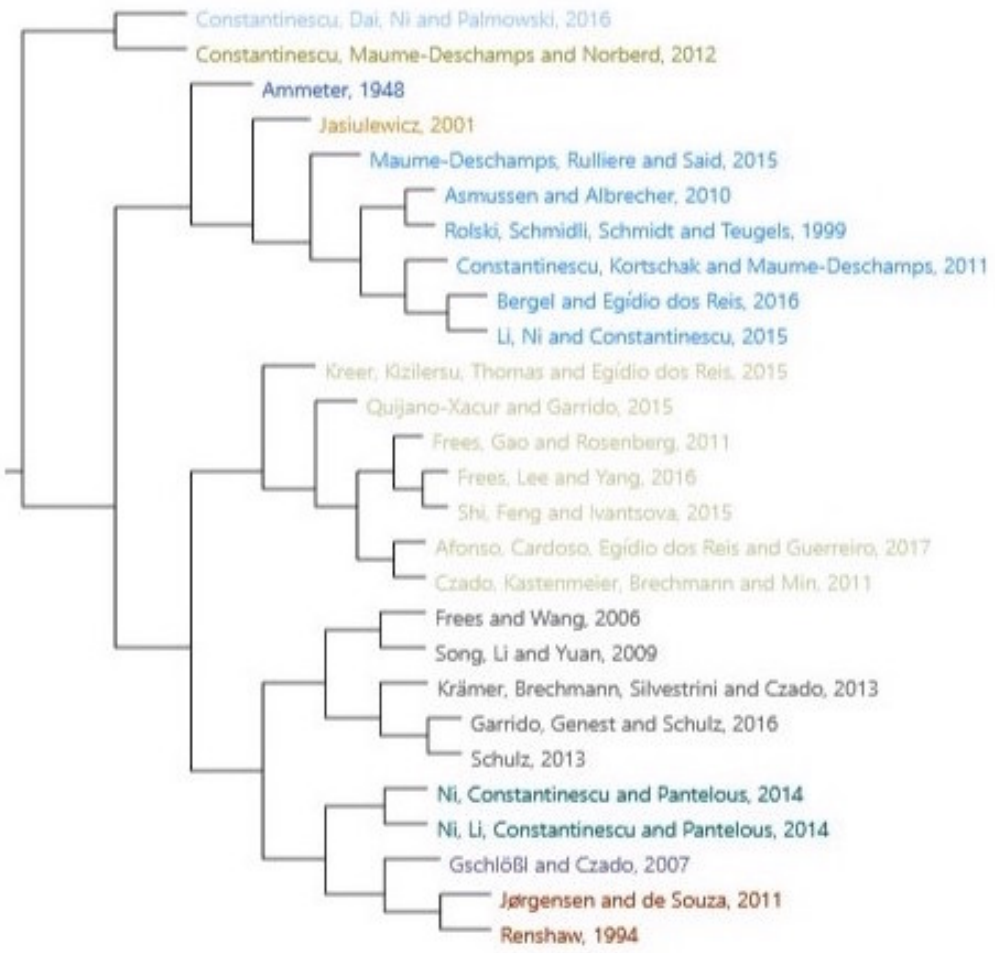

Figure 8: Cluster Analysis of the Sources - Dendrogram

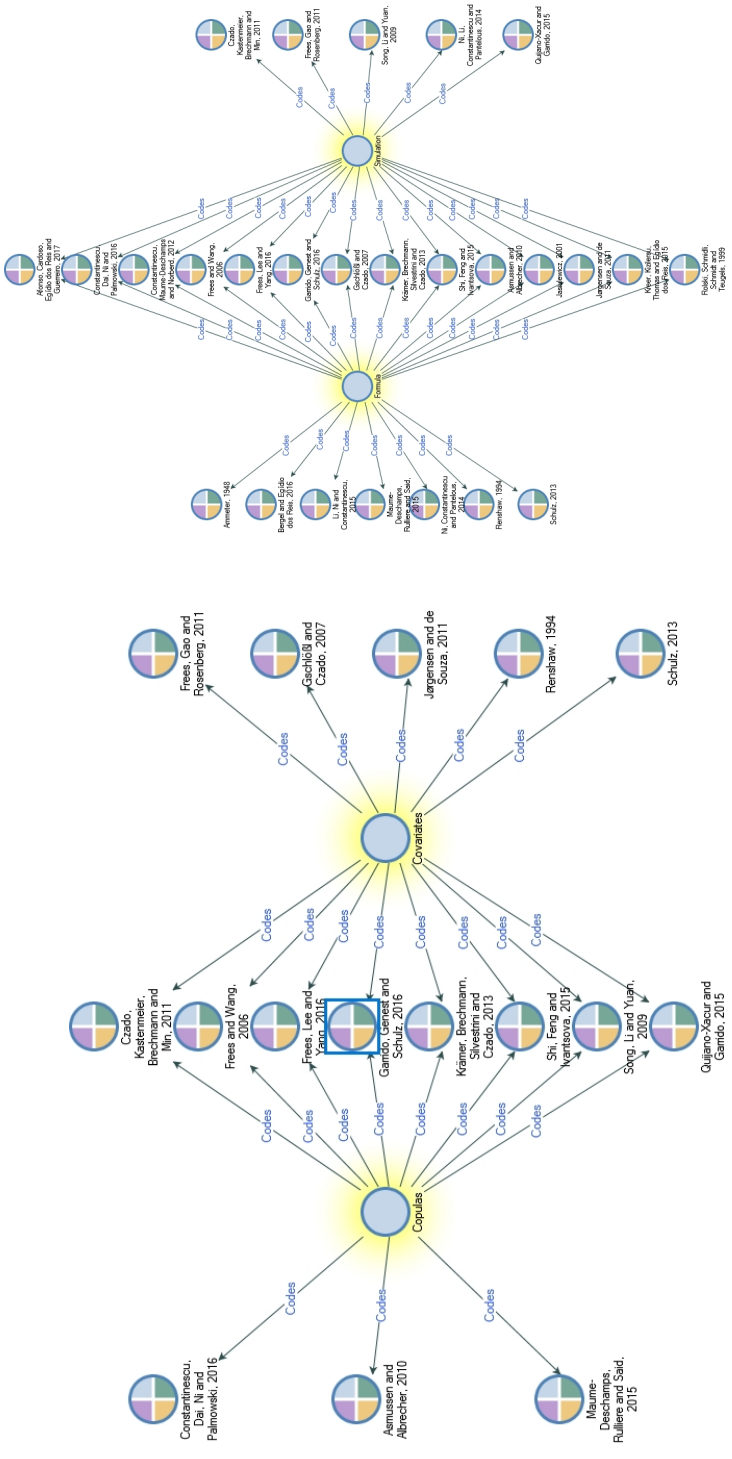

Figure 9: Comparison Diagrams

From the comparison diagram between "Copulas" and "Covariates" presented on the left of Figure 9, we can observe that from the 16 references coded for this two categories, Constantinescu *et al.* (2016), Asmussen and Albrecher (2010) and Maume-Deschamps *et al.* (2017) work only with copulas. Frees *et al.* (2011), Gschlößl and Czado (2007), Jørgensen and Paes De Souza (1994), Renshaw (1994) and Schulz (2013) use covariates on their papers. The others eight papers use them both. Both copulas and covariates are methods to try to capture the dependence between the claim frequency and severity variables.

A second comparison diagram is presented on the right of Figure 9, comparing "Formula" and "Simulation". We can point out that there are 26 papers, since those are the approach that the authors can follow to calculate the ruin probability or/and premium. So some authors used Formula, Ammeter (1948), Bergel and Egídio dos Reis (2016), Li *et al.* (2015), Maume-Deschamps *et al.* (2017), Ni *et al.* (2014b), Renshaw (1994) and Schulz (2013).

On the other hand the following authors from our database used "Simulation": Czado *et al.* (2011), Frees *et al.* (2011), Song *et al.* (2009), Ni *et al.* (2014a) and Quijano Xacur and Garrido (2015). The remaining authors used both. It's worth to comment that Ni, Constantinescu and Pantelous published two papers in 2014, one using "Formula" and the other "Simulation".

To finalize, we built the *framework matrix* where each row shows each paper and each columns the category mentioned above, in order to identify subtle connections which can allow a thorough and rigorous study. In Table 1 is presented a summarized version of this framework matrix, in which the first column presents the name of the cases in study, the following columns are 12 different categories and we mark the cells with an "×" to represent the coded categories to each source.

The categories presented in the Table 1 can be shortly defined as: *Actuarial/Actuaries:* Study of risk/ Scientist of risk; *Aggregate Claims Model:* Model of claims that considers both and all together frequency and severity of claims; *Claim Frequency:* Frequency or count of claims in the insurance company; *Claim Severity:* Severity or amount of claims in the insurance company.

Compound Poisson: Distribution for the aggregate claim amounts used to model the frequency and severity of claims on aggregate; *Copulas:* Is a multivariate probability tool used to capture dependence; *Joint Distribution:* Is the distribution of the two or more variables calculated together, jointly.

Premium: Amount paid by the insured for the insurance policy; *Regression:* Multiple Regression models, can also be GLM's; *Ruin Probability:* Probability of ruin of an insurance portfolio or company; *Simulation:* When simulating different scenarios on a software; *Stochastic Process:* Random Processes used for the claim frequency, in this case it's divided into Markov and Martingale processes.

Source	A	B	C	D	E	F	G	H	I	J	K	L
Afonso *et al.* (2017)			×	×	×			×		×	×	×
Ammeter (1948)	×							×				
Asmussen and Albrecher (2010)	×			×	×	×	×	×	×	×	×	×
Bergel and Egídio dos Reis (2016)				×	×					×		
Constantinescu *et al.* (2011)				×						×		×
Constantinescu *et al.* (2012)	×		×		×			×		×		×
Constantinescu *et al.* (2016)				×		×		×		×	×	×
Czado *et al.* (2011)	×		×		×	×	×	×	×		×	
Frees and Wang (2006)	×		×	×		×	×		×		×	×
Frees *et al.* (2011)	×						×		×		×	
Frees *et al.* (2016)	×	×	×	×		×	×	×	×		×	
Garrido *et al.* (2016)	×	×	×	×	×			×	×		×	
Gschlößl and Czado (2007)	×		×		×			×	×		×	×
Jasiulewicz (2001)	×							×		×		
Jørgensen and Paes De Souza (1994)				×	×		×	×	×		×	×
Krämer *et al.* (2013)	×		×	×		×	×		×		×	
Kreer *et al.* (2015)	×			×							×	
Li *et al.* (2015)					×			×		×		×
Maume-Deschamps *et al.* (2017)	×					×	×			×		
Ni *et al.* (2014a)			×	×				×				
Ni *et al.* (2014b)			×	×				×				
Quijano Xacur and Garrido (2015)			×	×	×			×	×		×	
Renshaw (1994)			×	×				×	×			
Rolski *et al.* (1999)	×			×	×		×	×	×	×	×	×
Schulz (2013)		×	×	×	×		×	×	×			
Shi *et al.* (2015)	×		×	×	×	×	×	×	×		×	×
Song *et al.* (2009)					×	×			×		×	

A: Actuarial; B: Aggregate Claims Model; C: Claim Frequency; D: Claim Severity; E: Compound Poisson; F: Copulas; G: Joint Distribution; H: Premium; I: Regression; J: Ruin Probability; K: Simulation; L: Stochastic Process

Table 1: Summarized Framework Matrix

4. Final Remarks

Our source intended to talk about the calculation of premiums and ruin probabilities for insurance application, also to associate the claim frequency with their severity. Some authors use copulas, others use covariates in a regression model, and others try to find a distribution that can capture that dependence.

We were motivated to organize and structure our research in Risk Theory and as presented in the paper, we were able to achieve this goal. And beyond that,

after a deeper study we extracted quantitative knowledge from the database.

We obtained results that made possible to know which authors were the most important for each category as we saw in Figure 3. It was shown in Figure 4 which categories matters the most for this data base and in which ways, hierarchically, the authors approach the subject.

Additionally, in Figure 5 we presented in percentage the relationship between the nodes. At last, from the cluster analysis shown in Figures 6, 7 and 8, we captured relevant patterns among the nodes and the authors.

The result showed to be interesting to compare respective categories and plot comparison diagrams, for instance, comparing *Dependence* with *Independence*; *Simulation* with *Formula*; *Copulas* with *Covariates*; *Regression* with *Copulas*; *Claim Severity* with *Claim Frequency* among others.

To finalize, this text mining analysis presents a current overview of the knowledge in the field of Ruin Theory research. In addition, a conceptual framework was presented and the key categories for the dependency model were identified. It is presumed that this study will motivate future research on the impact of dependence between these two variables on risk models, bringing to light the categories and links that need further investigation.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support from FCT - Fundao para a Cincia e a Tecnologia (Portuguese Foundation for Science and Technology) through Project CEMAPRE - UID/MULTI/00491/2013 financed by FCT/ MCTES through national funds.

References

Afonso, L. B., Cardoso, R. M. R., and Egídio dos Reis, A. D. (2017). Measuring the impact of a bonus-malus system in finite and continuous time ruin probabilities for large portfolios in motor insurance. *ASTIN Bulletin*, 47(2):417–435.

Aggarwal, C. C. and Zhai, C., editors (2012). *Mining Text Data*. Springer Science & Business Media, 1st edition.

Ammeter, H. (1948). A generalization of the collective theory of risk in regard to fluctuating basic-probabilities. *Scandinavian Actuarial Journal*, 1948(1-2):171–198.

Asmussen, S. and Albrecher, H. (2010). *Ruin Probabilities*. World Scientific, Singapore, second edition.

Bazeley, P. and Jackson, K. (2013). *Qualitative data analysis with Nvivo*. Sage Publications, second edition.

Bergel, A. I. and Egídio dos Reis, A. D. (2016). Ruin problems in the generalized Erlang(n) risk model. *European Actuarial Journal*, 6(1):257–275.

Chen, H., Fuller, S. S., Friedman, C., and Hersh, W. (2005). *Medical informatics: Knowledge management and Data Mining in biomedicine*, volume 8.

Cohen, M. A. and Hersh, R. W. (2005). A survey of current work in biomedical text mining. *Brief Bioninform*, 6(1):57 – 71.

Collier, N., Doan, S., Kawazoe, A., Goodwin, R. M., Conway, M., Tateno, Y., Ngo, Q. H., Dien, D., Kawtrakul, A., Takeuchi, K., Shigematsu, M., and Taniguchi, K. (2008). BioCaster: Detecting public health rumors with a Web-based text mining system. *Bioinformatics*, 24(24):2940–2941.

Constantinescu, C., Dai, S., Ni, W., and Palmowski, Z. (2016). Ruin probabilities with dependence on the number of claims within a fixed time window. *Risks*, 4(2):17.

Constantinescu, C., Kortschak, D., and Maume-Deschamps, V. (2011). Ruin probabilities in models with a Markov chain dependence structure. *Scandinavian Actuarial Journal*, 1238(December):1–24.

Constantinescu, C., Maume-Deschamps, V., and Norberg, R. (2012). Risk processes with dependence and premium adjusted to solvency targets. *European Actuarial Journal*, 2(1):1–20.

Corbin, J. and Strauss, A. (2008). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage Publications, USA, 3rd edition.

Corley, C. D., Cook, D. J., Mikler, A. R., and Singh, K. P. (2010). Text and structural data mining of influenza mentions in web and social media. *International Journal of Environmental Research and Public Health*, 7(2):596–615.

Czado, C., Kastenmeier, R., Brechmann, E. C., and Min, A. (2011). A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal*, 1238(January):1–28.

Feldman, R. and Sanger, J. (2006). *The Text Mining Handbook*. Cambridge University Press, New York, USA.

Francis, L. and Flynn, M. (2010). Text Mining Handbook. *Society*, (2008):1–61.

Frees, E., Lee, G., and Yang, L. (2016). Multivariate frequency-severity regression models in insurance. *Risks*, 4(1):4.

Frees, E. W., Gao, J., and Rosenberg, M. A. (2011). Predicting the frequency and amount of health care expenditures. *North American Actuarial Journal*, 15(3):377–392.

Frees, E. W. and Wang, P. (2006). Copula credibility for aggregate loss models. *Insurance: Mathematics and Economics*, 38(2):360–373.

Garrido, J., Genest, C., and Schulz, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics*, 70:205–215.

Gschlößl, S. and Czado, C. (2007). Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal*, 2007(3):202–225.

Han, J., Kamber, M., and Pei, J. (2012). *Data mining. Concepts and Techniques*. Elsevier, Waltham, USA, third edition.

He, W., Zha, S., and Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, 33:464–472.

Hirschman, L., Burns, G. A. P. C., Krallinger, M., Arighi, C., Cohen, K. B., Valencia, A., Wu, C. H., Chatr-Aryamontri, A., Dowell, K. G., Huala, E., Lourenço, A., Nash, R., Veuthey, A. L., Wiegers, T., and Winter, A. G. (2012). Text mining for the biocuration workflow. *Database*, 2012(November):1–10.

Jasiulewicz, H. (2001). Probability of ruin with variable premium rate in a Markovian environment. *Insurance: Mathematics and Economics*, 29(2):291–296.

Jørgensen, B. and Paes De Souza, M. C. (1994). Fitting Tweedie's compound poisson model to insurance claims data. *Scandinavian Actuarial Journal*, 1994(1):69–93.

Kim, J. M. and Jun, S. (2015). Graphical causal inference and copula regression model for apple keywords by text mining. *Advanced Engineering Informatics*, 29(4):918–929.

Krämer, N., Brechmann, E. C., Silvestrini, D., and Czado, C. (2013). Total loss estimation using copula-based regression models. *Insurance: Mathematics and Economics*, 53(3):829–839.

Kreer, M., Kizilersü, A., Thomas, A. W., and Egídio dos Reis, A. D. (2015). Goodness-of-fit tests and applications for left-truncated Weibull distributions to non-life insurance. *European Actuarial Journal*, 5(1):139–163.

Li, B., Ni, W., and Constantinescu, C. (2015). Risk models with premiums adjusted to claims number. *Insurance : Mathematics and Economics*, 65(2015):94–102.

Maume-Deschamps, V., Rullière, D., and Said, K. (2017). Impact of dependence on some multivariate risk indicators. *Methodology and Computing in Applied Probability*, 19(2):395–427.

Maynard, D., Bontcheva, K., and Rout, D. (2012). Challenges in developing opinion mining tools for social media. *Lrec 2012*, pages 15–22.

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications*, 40(10):4241–4251.

Ni, W., Constantinescu, C., and Pantelous, A. A. (2014a). Bonus-Malus systems with Weibull distributed claim severities. *Annals of Actuarial Science*, 8(02):217–233.

Ni, W., Li, B., Constantinescu, C., and Pantelous, A. A. (2014b). Bonus-Malus systems with hybrid claim severity distributions. *Vulnerability, Uncertainty, and Risk: Quantification, Mitigation, and Management*, pages 1234–1244.

Peng, T.-Q., Zhang, L., Zhong, Z.-J., and Zhu, J. J. (2012). Mapping the landscape of Internet Studies: Text mining of social science journal articles 2000-2009. *New Media & Society*, 15(5):644–664.

Quijano Xacur, O. A. and Garrido, J. (2015). Generalised linear models for aggregate claims: to Tweedie or not? *European Actuarial Journal*, 5(1):181–202.

Renshaw, A. E. (1994). Modelling the claims process in the presence of covariates. *ASTIN Bulletin*, 24(2):265–285.

Rolski, T., Schmidli, H., Schmidt, V., and Teugels, J. (1999). *Stochastic Processes for Insurance and Finance*. John Wiley & Sons, West Sussex, England.

Schulz, J. (2013). *Generalized Linear Models for a Dependent Aggregate Claims Model*. PhD thesis, Concordia University. Montréal, Quebec, Canada.

Shi, P., Feng, X., and Ivantsova, A. (2015). Dependent frequency-severity modeling of insurance claims. *Insurance: Mathematics and Economics*, 64:417–428.

Song, P. X.-K., Li, M., and Yuan, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics*, 65(1):60–68.

TechAmerica Foundation's Federal Big Data Commission (2012). Demystifying Big Data: A Practical Guide To Transforming The Business of Government. Technical report, TechAmerica Foundation's. Retrieved July 10, 2017, from https://www.attain.com/sites/default/files/take-aways-pdf/Solutions_Demystifying Big Data - A Practical Guide To Transforming The Business Of Government.pdf

Zeng, D., Chen, H., Lusch, R., and Li, S. H. (2010). Social media analytics and intelligence. *IEEE Intelligent Systems*, 25(6):13–16.

Zweigenbaum, P., Demner-fushman, D., Yu, H., and Cohen, K. B. (2007). Frontiers of biomedical text mining: Current progress. *Briefings in Bioinformatics*, 8(5):358–375.