Catastrophe Insurance: Estimation of the Generalized Tail Distortion Risk Measure and Earthquake and Wildfire Insurance Risk Modeling

Roba Bairakdar

A Thesis

In the Department of

Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy (Mathematics) at

Concordia University

Montréal, Québec, Canada

February 2023

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: Roba Bairakdar

Entitled: Catastrophe Insurance: Estimation of the Generalized Tail Distortion Risk Measure and Earthquake and Wildfire Insurance Risk Modeling

and submitted in partial fulfillment of the requirements for the degree of

Doctor Of Philosophy (Mathematics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

_____ Chair

Dr. T. Koreshkova

_____ External Examiner

Dr. J. Trufin

_____ Examiner

Dr. S. Brugiapaglia

_____ Examiner

Dr. Y. Lu

_____ Examiner

Dr. J. Schulz

_____ Supervisor

Dr. M. Mailhot

Approved by _____

Dr. Yogen Chaubey, Graduate Program Director

February 20, 2023 _____

Dr. Pascale Sicotte, Dean of Faculty

Abstract

Title: Catastrophe Insurance: Estimation of the Generalized Tail Distortion Risk Measure and Earthquake and Wildfire Insurance Risk Modeling

Roba Bairakdar, Ph.D.

Concordia University, 2023

In this thesis, we focus on catastrophic events in the context of insurance and risk management.

Insurance risk arising from catastrophes such as earthquakes is one of the components of the Minimum Capital Test for federally regulated property and casualty insurance companies. Given the spatial heterogeneity of earthquakes, the ability to assess whether the fits are adequate in certain locations is crucial in obtaining usable models. Accordingly, we extend the use of Voronoi residuals to calculate deviance Voronoi residuals. We also create a simulation-based approach, in which losses and insurance claim payments are calculated by relying on earthquake hazard maps of Canada. As an alternative to the current guidelines of OSFI, a formula to calculate the country-wide minimum capital test is proposed based on the correlation between the provinces. Finally, an interactive web application is provided which allows the user to simulate earthquake financial losses.

Homeowners' insurance in wildfire-prone areas can be a very risky business that some insurers may not be willing to undertake. We create an actuarial spatial model for the likelihood of wildfire occurrence over a fine grid map of North America. Several models are used, such as generalized linear models and tree-based machine learning algorithms. A detailed analysis and comparison of the models show a best fit using random forests. Sensitivity tests help in assessing the effect of future changes in the covariates of the model. A downscaling exercise is performed, focusing on some high-risk states and provinces. The

model provides the foundation for actuaries to price, reserve, and manage the financial risk from severe wildfires.

We explore the first and second-order asymptotic expansions of the generalized tail distortion risk measure for extreme risks. We propose to use the first-order asymptotic expansion to provide an estimator for this risk measure. The asymptotic normality of the estimator at intermediate and extreme confidence levels are shown, separately. Additionally, we provide bias-corrected estimators, where we focus on the case where the tail index γ is estimated by the Hill estimator. We perform a simulation study to assess the performances of the proposed estimators proposed and we compare them with other estimators in the literature. Finally, we showcase out estimator on several real-life actuarial data sets.

Acknowledgments

First and foremost, I would like to thank Allah for providing me with the strength and motivation that I required in this journey.

I would like to acknowledge and give my warmest thanks to my supervisor, Professor Mélina Mailhot, for her invaluable advice, trust, and support. I cannot thank her enough for all the knowledge she has bestowed upon me and for her enthusiasm, even during the toughest times. I am also extremely thankful to my collaborators, Professors Debbie Dupuis, Mathieu Boudreault, Frédéric Godin and Fan Yang. The opportunities I have had to work with a variety of actuaries and statisticians benefited my development and taught me immensely. The guidance of Professor Mailhot and our collaborators and their advice carried me through all the stages of my thesis.

My gratitude goes to my professors throughout my academic journey. I will forever be indebted to them, for they have been my teachers, guides, and inspiration.

A debt of gratitude is owed to the constructive feedback, meaningful comments and advice I received that helped shape my papers and set me on the right path for my future career.

I am thankful for the financial support of Concordia University, the Society of Actuaries' James C. Hickman Scholar program and the Casualty Actuarial Society.

Special thanks go to my loved ones. To my family and friends, for their love, support, and understanding. To my mom, for her continuous encouragement and trust. To Lecter, for keeping me sane when my mind told me otherwise. To Thakeb, for being my voice of reason. My words can never express how appreciative I am for all of you.

This thesis is dedicated to my grandmother and guardian angel, Samira, who passed away when I started my Ph.D. She was my biggest supporter and would be proud of my accomplishments. Mima, your memory will always be in my heart.

Contribution of Authors

This thesis is based on three research articles:

I : Bairakdar, R., F. Godin, M. Mailhot, and F. Yang. Estimation of the Generalized Tail Distortion Risk Measure.

This joint work is still in preparation and represents a draft. Bairakdar is responsible for a substantial portion of the Simulations and Application sections, as well as a portion of the writing with editing by Godin, Mailhot and Yang.

II : Bairakdar, R., D. Dupuis, and M. Mailhot. Goodness-of-fit tests for Space-time Point Process Models: An Application to Earthquake Insurance Risk.

This joint work is submitted to the Journal of Risk and Insurance. Bairakdar is responsible for a substantial portion of the analysis, as well as the primary portion of the writing with editing by Dupuis and Mailhot.

III : Bairakdar, R., B. Boudreault, and M. Mailhot. Random Forests for Wildfire Insurance Applications.

This joint work is accepted for publication in Variance Journal. Bairakdar is responsible for a substantial portion of the analysis, as well as the primary portion of the writing with editing by Boudreault and Mailhot.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Catastrophic events are events that cause great and unexpected damage. In the insurance context, they are categorized as events that have low probability of occurrence and high claim amounts. Managing catastrophic risks includes identifying and quantifying the insurance risk from such events, pricing and underwriting insurance products based on the risk profile of the insureds, deciding on the risk tolerance level of the organization, diversifying the portfolio, sharing the risk with reinsurers, and setting sufficient capital or reserves to ensure the company's readiness against such extreme events. A thorough and complete understanding of the insurance risk associated with catastrophes guarantees that the charged premiums are adequate and that the reinsurance contracts and capital are sufficient, and hence minimizes the risks of insufficient funds to cover possible risks. Catastrophic events may include natural disasters such as earthquakes, wildfires ignited due to lightnings, floods, and hurricanes and also man-made disasters such as terrorist attacks, riots, and wildfires ignited by human-kind. Such events may be excluded from a standard homeowners insurance policy, depending on the regulation and policy conditions set by each insurer and reinsurer. Typically, the insurer sets its definition of a catastrophe based on an event exceeding at least one of several thresholds for loss, such as the total insured loss amount or the number of casualties, etc. For example, Canada's

loss and exposure indices provider, Catastrophe Indices and Quantification Inc. (CatIQ), define a catastrophe as an event that causes more than CAD $25m of insured damage (CatIQ 2021). Since 2014, CatIQ has collected over 130 events in their database, corresponding to CAD $21b of insured losses.

1.1 Insurance Risk Models

Predictive modeling for property and casualty (P&C) insurance companies usually requires considering the frequency and severity of the claims, separately. Pricing actuaries traditionally used generalized linear models (GLMs) to develop insurance pricing models; see, for example, De Jong, Heller, et al. (2008), Ohlsson and Johansson (2010), Anderson et al. (2004), and Haberman and Renshaw (1996). In recent years, machine learning algorithms are becoming more popular, and accordingly predictive modeling is gradually changing. Machine learning techniques in the actuarial literature are still limited. Wüthrich (2018) and Wüthrich and Buser (2020) use tree-based algorithms to model claim frequencies, while Schelldorfer and Wüthrich (2019) use neural networks. Guelman (2012) uses gradient boosting trees for modeling and predicting auto insurance loss costs. Henckaerts et al. (2021) use decision trees to develop pricing plans built from the frequency and severity of claims. Gabrielli, Richman, and Wüthrich (2020) embed a classical actuarial regression model into a neural network architecture for reserving purposes.

In the previously mentioned models, which cover non-catastrophic insurance products, the actuaries are responsible of measuring the risk by relying on the historical claim experience of the insurance company. This requires the availability of enough data points to capture a wide range of possible probabilistic scenarios, in addition to proper on-levelling of the claims data, i.e. manipulating the data to ensure that the historical risk is comparable to

the current level of risk, by for example accounting for inflation, or an increase in exposure. This method could be difficult for catastrophe risk due to the unavailability of a sufficient number of data points because, by definition, catastrophes are extreme events and they have very long return periods, i.e. low probability of occurrence and high claim amounts. Accordingly, losses in recent years may not be indicative of future losses. It is highly improbable that an insurance company at the current time has a complete database of extreme losses, such as the loss that occurs at a rate of 1-in-500 years, and even if they have 500 years worth of data, using the data could be impractical due to the possible changes that could have happened in 500 years, such as climate change, urbanization, building codes, building material, etc.

Accordingly, it is more appropriate to consider a risk decomposition approach for catastrophe modeling, i.e. identify the hazard, vulnerability and exposure; see Mitchell-Wallace et al. (2017) for details. There is catastrophe risk at the intersection of those three components, such that a vulnerable property is subject to wildfire risk, for example, due to its location and characteristics. Firstly, the hazard portion of the model indicates the intensity of the risk, such as the frequency or rate of wildfire occurrence or peak ground acceleration for an earthquake, which are both examples of a hazard that varies spatially. Secondly, vulnerability assessment helps in quantifying the relationship between the hazard and the damage by means of a metric, such as damage ratios, which quantifies the expected damage as a percentage of the replacement cost of the building. This provides a tool to estimate the relative damage given the intensity of the hazard. For catastrophic events, the lack of sufficient data points may pose a hurdle in the creation of vulnerability functions and relations. In these cases, the help of civil engineers may be required to analyze the characteristics of the properties and estimate the vulnerability relations to the hazard in question. Thirdly, collection of exposure requires the valuation of properties and/or infrastructures at risk. This is usually represented in terms of the cost of building replacement. The values are usually divided into buildings

and contents exposure, and can be divided further into building occupational classes, such as residential and non-residential, and building construction type, such as wooden, concrete, steel, masonry, etc, and year of built. In earthquake modeling, for example, such information are major factors that explain the extent of damage caused by an earthquake and its subsequent events. For example, an earthquake in Haïti in 2010 resulted in almost 250,000 fatalities, however an earthquake in New Zealand in 2010, of comparable intensity, did not cause any fatalities. A result attributed to New Zealand's advanced building code and readiness to earthquakes (Wallemacq and House 2017). Industry exposure database could also be outsourced from external providers, such as CatIQ. Finally, a typical catastrophe model uses the previous components to calculate the total losses at the predetermined aggregation level, in addition to claim values, which are calculated after the application of insurance policy terms, such as deductibles and policy limits. Very small claim amounts that are associated with internal expenses may be obsolete for an insurance company to cover because the price of the policy will be unreasonable. Accordingly, some insurance policies may have a deductible, which is the amount that the insured will have to cover before getting a claim payment from the insurer. For very large claim values, the insurer may wish to cover the damage up to a predetermined policy limit. Deductibles and limits can vary by coverage and can be set as fixed values, or a proportion of the exposure or the total loss.

Given that catastrophes are usually geographically correlated risks, assuming complete independence between the individual risks is unrealistic in practice. For example, consider a wildfire flaming in a highly populated area. It is highly probable that neighboring properties may experience damage from that wildfire at the same time, which means that one could have spatial dependence in between the risks in this event. Conditional on a set of spatially dependent variables, one could achieve spatial independence. It is also realistic to assume spatial heterogeneity, which means that observations are not identically distributed. Spatial modeling in the actuarial literature

usually focuses on claims data; see, for example Shi and Shi (2017) for spatial frequency and severity modeling of auto-insurance claims. However, as discussed earlier, one may not always have sufficient claims data that covers the geographical region of interest, especially for natural catastrophes. Accordingly, a catastrophe decomposition approach is to be pursued, as we demonstrate in Chapter 4.

Spatio-Temporal Point Processes (STPP) can be used to model data of the location and time of occurrence of events. Assessment of the adequacy of the fit of STPP models has witnessed considerable progress in recent years. For example, Schorlemmer and Gerstenberger (2007) developed methods such as the Number test, which compares the total estimated and observed count of events with the observed, the Likelihood test, which assesses the overall likelihood of the fitted model, and the Likelihood-ratio test, which compares the relative performance of two fitted models, see Schorlemmer and Gerstenberger (2007), Zechar, Gerstenberger, and Rhoades (2010), and Rhoades et al. (2011). Those methods are successful in assessing the overall fit of the model, however, they are incapable of identifying locations or time periods where the fit is poor. Recent progress in measuring the goodness-of-fit and comparing spatio-temporal models include pixel-based residual methods, in which the spatial region is divided into a predetermined regular grid and the residual is computed for each pixel, see Baddeley et al. (2005) and Zhuang (2006). For models where the expected number of events in a pixel is close to zero, some problems may arise because of the heavy skewness of the distribution of the residual for the given pixel. Bray et al. (2014) proposed to compare the expected and observed count of events over Voronoi polygons generated by the observed spatio-temporal point pattern.

1.2 Risk Measures

Risk management is essential for regulators, insurance companies and financial institutions for several reasons, such as pricing insurance products and assessing the risk exposure of the entity to determine the minimum required capital that needs to be held to cover unexpected losses. For example, the Office of the Superintendent of Financial Institutions (OSFI) in Canada sets guidelines for the recommended assessment of insurers' catastrophe risk management and calculation of probable maximum loss (PML), the threshold dollar value of losses beyond which losses are unlikely. The insurance risk arising from earthquakes is one of the components of the Minimum Capital Test (MCT) for federally regulated P&C insurance companies and it is calculated in terms of the PML that occurs at a rate of 1-in-500 years, representing the 99.8^{th} percentile of the distribution of the annual maxima.

Risk measures can also be used in pricing insurance products and in capital allocation per line of business. It helps the organizations manage the risks that emerge from adverse scenarios, especially those that are highly unlikely. They are typically used to explain the tail behavior of the loss distributions. Various risk measures have been proposed in the literature, such as Value-at-Risk (VaR), Tail Value-at-Risk (TVaR), Range Value-at-Risk (RVaR), distortion risk measure, etc. VaR is the most commonly known risk measure, and it represents the maximum loss amount which is not exceeded with a given high confidence level p. By definition, the VaR does not provide any information about the behavior of the loss random variable X beyond its p-level quantile, which is one of the disadvantages of using VaR as a risk measure. Additionally, it lacks the subadditivity property for coherent risk measures; see Artzner et al. (1999) and McNeil, Frey, and Embrechts (2015) for further details and properties of risk measures. Thus, TVaR was proposed as an alternative to VaR and it is usually superior because it is subadditive and it considers losses beyond the VaR. For a continuous random variable, TVaR is simply defined as the expected loss in the event that VaR at confidence level p is exceeded. TVaR is sometimes referred to as

the Expected Shortfall. RVaR is a general case that combines both VaR and TVaR. It is defined as the expected loss in the event that VaR at confidence level p_1 is exceeded, but still below VaR at confidence level p_2, for $p_1 \leq p_2$.

Definition 1.2.1. *For a continuous random variable X with cumulative distribution function F_X, the Range Value-at-Risk at level range (p_1, p_2), with $0 \leq p_1 \leq p_2 \leq 1$ is defined as*

$$\mathrm{RVaR}_{p_1,p_2}(X) = \mathbb{E}\left[X \middle| \mathrm{VaR}_{p_1}(X) \leq X \leq \mathrm{VaR}_{p_2}(X)\right] = \begin{cases} \frac{1}{p_2-p_1}\int_{p_1}^{p_2} \mathrm{VaR}_u(X)\mathrm{d}u, & \text{if } p_1 < p_2 \\ \mathrm{VaR}_{p_1}(X), & \text{if } p_1 = p_2, \end{cases}$$

where

$$\mathrm{VaR}_p(X) = \inf\left\{x \in \mathbb{R} : F_X(x) \geq p\right\}$$

is the Value-at-Risk at level $p \in [0,1]$.

For $p_1 = p_2$, RVaR simplifies to VaR. For $p_1 < p_2 < 1$, RVaR is well defined for any random variable. For $p_1 < p_2 = 1$, RVaR simplifies to TVaR, which is well defined only for the set of random variables with finite means. Even though VaR fails to give any information beyond the level p, RVaR quantifies the magnitude of the loss of the worst $100(1-p_1)\%$ to $100(1-p_2)\%$ cases. VaR and TVaR can be expressed in terms of Choquet Integrals (Denneberg 1994).

Definition 1.2.2. *Let $g(\cdot) : [0,1] \to [0,1]$ be a non-decreasing function with $g(0) = 0$ and $g(1) = 1$. For a non-negative continuous random variable X with cumulative distribution function F_X and survival function $\overline{F}_X$, the Choquet integral is defined as follows:*

$$\rho_g(X) = \int_0^\infty g\left(\overline{F}_X(x)\right)\mathrm{d}x. \tag{1.2.1}$$

This function is a distortion function, and given that X is a loss random variable, then it is also called a distortion risk measure.

Let $g_{\mathrm{VaR}}(x) = \mathbb{1}_{x>1-p}$ and $g_{\mathrm{TVaR}}(x) = \min\{x/(1-p), 1\}$, then it can be easily shown that

$$\mathrm{VaR}_p(X) = \int_0^\infty g_{\mathrm{VaR}}\left(\overline{F}_X(x)\right) \mathrm{d}x$$
$$\mathrm{TVaR}_p(X) = \int_0^\infty g_{\mathrm{TVaR}}\left(\overline{F}_X(x)\right) \mathrm{d}x.$$

Denneberg (1990) proposed using distortion functions as premium principles and they have been developed further afterwards, see for example S. S. Wang (2000). Denote by q the quantile function of X. By applying some change of variables and integration by parts, an alternative representation of Eq. (1.2.1) can be provided,

$$\rho_g(X) = \int_0^1 g(u)\mathrm{d}q(1-u) = \int_0^1 q(1-u)\mathrm{d}g(u).$$

To study the asymptotic behavior of distortion risk measures, Zhu and Li (2012) introduced tail distortion risk measures, which represent the expected losses under a scenario probability measure that is deformed on the tail loss distribution.

Definition 1.2.3. *Let $g(\cdot) : [0,1] \to [0,1]$ be a non-decreasing function with $g(0) = 0$ and $g(1) = 1$. For a non-negative continuous random variable X, the tail distortion risk measure is defined as follows:*

$$\rho_g(X|X > \mathrm{VaR}_p(X)) = \rho_{g,p}(X) = \int_0^\infty g\left(\overline{F}_{X|X>\mathrm{VaR}_p(X)}(x)\right) \mathrm{d}x, \tag{1.2.2}$$

where $\overline{F}_{X|X>t}(x) = P(X > x|X > t)$.

With a distortion function $g(\cdot)$, introduce $g_p(\cdot)$ with $0 < p < 1$ as

$$g_p(u) = \min\{u/(1-p), 1\} = \begin{cases} g\left(\frac{u}{1-p}\right), & 0 \le u < 1-p, \\ 1, & 1-p \le u \le 1, \end{cases} \tag{1.2.3}$$

which again is a distortion function. By combining Eq. (1.2.2) and Eq. (1.2.3), the tail distortion risk measure at confidence level p of a loss random variable X can be re-written

as

$$\rho_{g,p}(X) = \int_0^\infty g_p\left(\overline{F}(x)\right) \mathrm{d}x.$$

For the special case where $g(x) = x$, the tail distortion risk measure becomes TVaR. El Methni and Stupfler (2017) propose a more general risk, called the extreme Wang distortion risk measure.

$$\rho_{g,p}(h(X)) = \int_0^1 h \circ q(1-u)\mathrm{d}g_p(u).$$

where $h : [0,\infty) \to [0,\infty)$ is a strictly increasing, continuously differentiable function. This generalized tail distortion risk measure covers many well known risk measures. For example, if $g(x) = x$ and $h(x) = x^\alpha$, then

$$\rho_{g,p}(X) = E\left[X^\alpha | X > \mathrm{VaR}_p(X)\right],$$

which is the Conditional Tail Moment (CTM) of order α of the loss random variable X.

1.3 Contributions

In Chapter 2, we explore the first and second-order asymptotic expansions of the generalized tail distortion risk measure for extreme risks. We propose to use the first-order asymptotic expansion to provide an estimator for this risk measure. El Methni and Stupfler (2017) had proposed estimator based on the asymptotic expansions but under the assumption that $h(x) = x^\alpha$. Alternatively, we give a complete treatment of the estimator for general h. We also prove the asymptotic normality of the estimator at intermediate and extreme confidence levels, separately. Additionally, we provide bias-corrected estimators, where we focus on the case where the tail index γ is estimated by the Hill estimator. We perform

a simulation study to assess the performances of the proposed estimators proposed and we compare them with other estimators in the literature, such as the ones introduced in El Methni and Stupfler (2017, 2018). Finally, we showcase out estimator on several real-life actuarial data sets.

In Chapter 3, we prepare a catastrophe model for earthquakes, in which the distribution of losses and insurance claim payments are estimated for a set of simulated earthquakes. We also analyze the earthquake insurance risk in Canada by creating an open-source and reproducible simulation-based approach. We start by fitting multiple STPP models, which require testing to confirm their adequacy and model comparison techniques to choose the best fit model. We propose to extend the Voronoi residual methods by calculating the Pearson Voronoi residuals, which have mean 0 and variance approximately equal to 1, in analogy with Pearson residuals for generalized linear models. For comparison of competing models, we also propose to extend the definition of pixel-based deviance residuals to Voronoi Polygons. In analogy with linear models, the resulting residuals may be called Deviance Voronoi residuals. Those methods are used to measure the goodness-of-fit of fitted spatio-temporal models for the significant Canadian earthquakes point pattern. The best fit model is used to simulate a large number of earthquakes. The simulation algorithm relies on publicly available data, such as earthquake hazard maps of Canada and population census data. For each earthquake, we estimate its location, magnitude, intensity and the affected properties. Building occupancy classifications and their respective damage probability matrices are combined with the earthquake insurance market penetration and policy terms to calculate the earthquake risk for each Canadian municipality. This models allows us to extend earthquake losses beyond past events into probable occurrences in the future and provide a methodology to quantify the insured losses from earthquake occurrences. An interactive web application allowing the user to simulate a significant earthquake based on any chosen location is provided. Additionally, we review OSFI's MCT formula for earthquake insurance risk in P&C companies and provide a possible alternative.

In Chapter 4, we create an actuarial spatial model for the likelihood of wildfire occurrence over a fine grid map of North America. The wildfire risk model is specifically tailored for USA and Canada, separately, and is split for each cause of ignition: man-made and natural. The model takes into account multiple types of wildfire risk features: continuous variables, categorical variables, and geographical coordinates. Several models are used, such as generalized linear models and tree-based machine learning algorithms, see Friedman, Hastie, Tibshirani, et al. (2001). Combining numerous decision trees in an ensemble, such as a random forest, provides high predictive performance with an ability to discover interaction effects between the predictors. Tree-based machine learning models offer strong predictive capabilities, in addition to their interpretability and ease of explanation of the importance of the predictors, unlike neural networks. They can also implicitly handle variable correlation, capture variable interactions and non-linear relationships between the predictors. Given the lack of publicly available wildfire claims data for each location in North America, we instead focus on modeling the annual burn probabilities by relying on publicly-available historical wildfire data. The goal of the article is to provide an interpretable model that predicts the annual burn probabilities for any location in North America and to compare how random forest models perform compared to more commonly used actuarial modeling techniques, such as GLMs. The comparison focuses on the statistical accuracy and interpretability of the models. The best fit model is used in a downscaling exercise where we predict the annual burn probabilities at a high resolution for some high-risk statues and provinces. The model provides the foundation for actuaries to price, reserve, and manage the financial risk from severe wildfires.

At the end of the thesis, we provide an appendix that supplements the chapters of the thesis. It contains some definitions, additional details, algorithms and results that are mentioned within the text. Throughout the text, references are given to alert the reader if they wish to visit the appendix.

Chapter 2

Estimation of the Generalized Tail Distortion Risk Measure

2.1 Introduction

Risk measures are now well established statistics, which are used in several fields such as environment, biomedical, actuarial science and enterprise risk management. The objective is always to represent specific characteristics of a risk, respecting the users' preferences and risk appetite. The risk measure for a loss random variable is the amount of additional capital requirement that a financial institution needs to invest as a buffer against unexpected future losses, such that the resulting position is satisfactory to internal or external risk parties. Distortion risk measures represent expected losses under a scenario probability measure, which may represent a stakeholder's preference. They were first explored in the actuarial literature by Denneberg (1994) and S. Wang (1995). Denneberg (1990) proposed using distortion functions as premium principles and they have been developed further afterwards, see for example S. S. Wang (2000).

Definition 2.1.1. *Let $g(\cdot) : [0,1] \to [0,1]$ be a distortion function, namely a non-decreasing function with $g(0) = 0$ and $g(1) = 1$. For a random variable X with cumulative distribution*

function F_X and survival function $\overline{F}_X$, the distortion risk measure of X is defined as follows:

$$\rho_g(X) = \int_{-\infty}^{0} \left[g\left(\overline{F}_X(x)\right) - 1 \right] \mathrm{d}x + \int_{0}^{\infty} g\left(\overline{F}_X(x)\right) \mathrm{d}x.$$

For a non-negative and continuous risk variable X, the first term in the equation above cancels, and hence it coincides with Definition 1.2.2.

Some of the most common risk measures are in fact special cases of distortion risk measures. For example, the Value-at-Risk (VaR) and the Tail Value-at-Risk (TVaR) of a non-negative loss random variable can be expressed as distortion risk measures for $g(x) = \mathbb{1}_{\{x>1-p\}}$ and $g(x) = \min\{x/(1-p), 1\}$, respectively. The distortion risk measure ρ_g satisfies several properties, which are listed below.

Lemma 2.1.1. *Let $g(\cdot) : [0,1] \to [0,1]$ be a distortion function, and X be a continuous random variable. Then the distortion risk measure ρ_g satisfies the following properties:*

1. *(Translation invariance) Consider the scalar $c \in \mathbb{R}$. Then $\rho_g(X + c) = \rho_g(X) + c$.*
2. *(Positive homogeneity) Consider the scalar $c \in \mathbb{R}$. Then $\rho_g(cX) = c\rho_g(X)$.*
3. *(Monotonicity) Consider the continuous random variable Y, such that $X \leq Y$ almost surely. Then $\rho_g(X) \leq \rho_g(Y)$.*
4. *(Subadditivity) Consider the scalars $c, x, y \in [0,1]$, such that $g\left(cx + (1-c)y\right) \geq cg(x) + (1-c)g(y)$, i.e. g is concave, and let Y be a continuous random variable. Then $\rho_g(X + Y) \leq \rho_g(X) + \rho_g(Y)$.*
5. *(Superadditivity) Consider the scalars $c, x, y \in [0,1]$, such that $g\left(cx + (1-c)y\right) \leq cg(x) + (1-c)g(y)$, i.e. g is convex, and let Y be a continuous random variable. Then $\rho_g(X + Y) \geq \rho_g(X) + \rho_g(Y)$.*

Note that a risk measure is coherent if it satisfies properties 1-4 in Lemma 2.1.1 (Artzner et al. 1999) and not all distortion risk measures are coherent, such as the VaR, which does

not satisfy the subadditivity property. The relationship between coherent and distortion risk measures were introduced by Schmeidler (1986, 1989).

Zhu and Li (2012) study the asymptotic behavior of distortion risk measures focusing on the right tail of a heavy-tailed loss distribution. They introduce the tail distortion risk measure, which represents the expected losses under a scenario probability measure that is deformed on the tail loss distribution.

Definition 2.1.2. *With a distortion function $g(\cdot)$, introduce $g_p(\cdot)$ with $0 < p < 1$ as*

$$g_p(u) = \begin{cases} g\left(\frac{u}{1-p}\right), & 0 \leq u < 1-p, \\ 1, & 1-p \leq u \leq 1, \end{cases}$$

which again is a distortion function. Then for a random variable X with cumulative distribution function F_X and survival function $\overline{F}_X$, the tail distortion risk measure of X at level p is defined as follows:

$$T_p[X] = \int_{-\infty}^{0} \left[g_p\left(\overline{F}(x)\right) - 1\right] \mathrm{d}x + \int_{0}^{\infty} g_p\left(\overline{F}(x)\right) \mathrm{d}x$$

This definition differs from the one given by Zhu and Li (2012), but they are identical when the risk variable X is non-negative and continuous, as it is expected to be in the context of insurance.

Clearly, if $g(\cdot)$ is concave, then $g_p(\cdot)$ is concave as well, which leads to the coherence of the tail distortion risk measure. For the special case with $g(x) = x$ and a continuous risk variable X, for $p > F(0)$ the tail distortion risk measure becomes the well-known the expected shortfall, $T_p[X] = \mathrm{E}\left[X | X > F^{\leftarrow}(p)\right]$, where, and throughout the paper,

$$F^{\leftarrow}(p) = \inf\{x : F(x) \geq p\}$$

is the VaR of X or the quantile of F with the usual convention that $\inf \varnothing = \infty$.

Define the function $U(\cdot)$ as the quantile function of $1/\overline{F}$, namely,

$$U(t) = \left(\frac{1}{\overline{F}}\right)^{\leftarrow}(t) = F^{\leftarrow}\left(1 - \frac{1}{t}\right), \qquad t \geq 1.$$

We assume $g(\cdot)$ and $g_p(\cdot)$ to be left continuous, then we can rewrite $T_p[X]$ as

$$T_p[X] = \int_0^1 F^{\leftarrow}(1-q)\mathrm{d}g_p(q) = \int_0^1 U\left(\frac{1}{q(1-p)}\right)\mathrm{d}g(q). \tag{2.1.1}$$

In El Methni and Stupfler (2017), a more general risk, called extreme Wang distortion risk measure, is proposed

$$\rho_p[X] = \int_0^1 h \circ U\left(\frac{1}{q(1-p)}\right)\mathrm{d}g(q) \tag{2.1.2}$$

where $h : [0,\infty) \to [0,\infty)$ is a strictly increasing, continuously differentiable function. Various estimation methods of $\rho_p[X]$ are studied by El Methni and Stupfler (2017, 2018).

This generalized tail distortion risk measure covers many well known risk measure. For example, if $g(x) = x$ and $h(x) = x^\alpha$, then

$$\rho_p(X) = E\left[X^\alpha | X > F^{\leftarrow}(p)\right],$$

known as the Conditional Tail Moment (CTM) of order α of the loss random variable X.

This article is divided as follows. In Section 2.2 we explore the first and second-order asymptotic expansions of the generalized tail distortion risk measure for extreme risks. In Section 2.3, we provide estimators for the generalized tail distortion risk measures and we prove their asymptotic normality at intermediate and extreme confidence levels p, separately. El Methni and Stupfler (2017) had proposed estimator based on the asymptotic expansions but under the assumption that $h(x) = x^\alpha$. Alternatively, we give a complete treatment of the estimator for general h. Additionally, we provide bias-corrected estimators, where we focus on the case where the tail index γ is estimated

by the Hill estimator. In Section 2.4, we perform a simulation study to assess the performances of the proposed estimators proposed and we compare them with other estimators in the literature, such as the ones introduced in El Methni and Stupfler (2017, 2018). In Section 2.6, we showcase our estimator on several real-life actuarial data sets. Finally, Section 2.7 concludes the article.

2.2 Asymptotic Analysis

In this section, we explore the asymptotic expansions of the generalized tail distortion risk measure for extreme risks. For a random variable X with distribution function F, we say that X is regularly varying with extreme value index $\gamma > 0$ if its survival function, $\overline{F}$, is regularly varying with index $-1/\gamma$. This is also denoted by $X \in \mathrm{RV}_{-1/\gamma}$, where γ index follows from Definition 2.2.1.

Definition 2.2.1. *An eventually non-negative measurable function $f(\cdot)$ is said to be* regularly varying *at ∞ with index $\gamma \in \mathbb{R}$, if for all $x > 0$,*

$$\lim_{t\to\infty} \frac{f(tx)}{f(t)} = x^{\gamma}. \tag{2.2.1}$$

This is denoted by $f(\cdot) \in \mathrm{RV}_{\gamma}$.

The second-order regular variation is needed for analyzing the asymptotic normality of the estimator.

Definition 2.2.2. *A regularly varying function $f(\cdot)$ is said to be* second-order regularly varying *at ∞ with first-order index $\gamma \in \mathbb{R}$ and second-order index $\rho \leq 0$, if there exists a measurable function $A(\cdot)$, which does not change sign eventually and converges to 0, such that*

$$\lim_{t\to\infty} \frac{f(tx)/f(t) - x^{\gamma}}{A(t)} = x^{\gamma}\frac{x^{\rho}-1}{\rho} =: J_{\gamma,\rho}(x). \tag{2.2.2}$$

This is denoted by $f(\cdot) \in 2\mathrm{RV}_{\gamma,\rho}$ and A is called (second-order) auxiliary function.

When $\rho = 0$, $J_{\gamma,\rho}(x)$ is understood as $x^\gamma \log x$.

When the survival function $\overline{F}$ is assumed to be RV or 2RV, it is equivalent to assume that the tail quantile function U is RV or 2RV. To be more specific, Theorem 1.1.6 and Corollary 1.2.10 of Haan and Ferreira (2006) show that for $\gamma < 0$, $U(\cdot) \in \mathrm{RV}_\gamma$ if and only if $\overline{F}(\cdot) \in \mathrm{RV}_{-1/\gamma}$. Further, Theorem 2.3.9 of Haan and Ferreira (2006) shows that for $\gamma > 0$ and $\rho \leq 0$, $U(\cdot) \in 2\mathrm{RV}_{\gamma,\rho}$ with an auxiliary function $A(\cdot)$ if and only if $\overline{F}(\cdot) \in 2\mathrm{RV}_{-1/\gamma,\rho/\gamma}$ with an auxiliary function $A(1/\overline{F}(\cdot))$. In this case, necessarily $A(\cdot) \in \mathrm{RV}_\rho$.

Let a random variable Y have the tail quantile function $V(t) = h \circ U(t)$ and denote its distribution function by G. Then by the definition of generalized tail distortion risk measure in Eq. (2.1.2), it is equivalent to have

$$\rho_p(X) = \int_0^1 V\left(\frac{1}{q(1-p)}\right) dg(q) = T_p(Y).$$

That is the generalized tail distortion risk measure of X is equivalent to the tail distortion risk measure of Y. Then with the following regular variation result on the tail quantile function V, the asymptotic expansions of $\rho_p(X)$ can be obtained directly from Yang (2015).

Lemma 2.2.1. *1. (Lemma 2.4 (i) of Lv, Mao, and Hu (2012)) If $U \in \mathrm{RV}_\gamma$ with $\gamma > 0$ and $h \in \mathrm{RV}_\alpha$ with $\alpha > 0$, then $V = h \circ U \in \mathrm{RV}_\xi$ with $\xi = \alpha\gamma$.*

2. (Proposition 2.9 of Lv, Mao, and Hu (2012)) Assume that $U \in 2\mathrm{RV}_{\gamma,\rho}$ with $\gamma > 0$, $\rho < 0$, and auxiliary function $A(t)$, and $h \in 2\mathrm{RV}_{\alpha,\beta}$ with $\alpha > 0$, $\beta < 0$ and auxiliary function $B(t)$. Then $V = h \circ U \in 2\mathrm{RV}_{\xi,\eta}$ with auxiliary function $C(t)$, where $\xi = \alpha\gamma$, $\eta = \max\{\rho, \beta\gamma\}$, and $C(t) = \alpha A(t) + \gamma B \circ U(t)$.

2.2.1 First-order Asymptotics

Now we are ready to show the first-order asymptotic expansions of the generalized tail distortion risk measure, which is by Corollary 3.1(a) of Yang (2015).

Theorem 2.2.1. *Assume that* $U \in \mathrm{RV}_\gamma$ *with* $\gamma > 0$ *and* $h \in \mathrm{RV}_\alpha$ *with* $\alpha > 0$. *Further assume that* $\int_1^\infty g\left(x^{-1/(\alpha\gamma+\delta)}\right) \mathrm{d}x < \infty$ *for some* $\delta > 0$. *Then*

$$\rho_p[X] \sim \lambda_\gamma h\left(F^{\leftarrow}(p)\right), \tag{2.2.3}$$

where $\lambda_\gamma = 1 + \int_1^\infty g\left(x^{-1/(\alpha\gamma)}\right) \mathrm{d}x$.

2.2.2 Second-order Asymptotics

Next, we show the second-order asymptotic expansions of the generalized tail distortion risk measure, which is by Corollary 4.1(a) of Yang (2015).

Theorem 2.2.2. *Assume that* $U \in 2\mathrm{RV}_{\gamma,\rho}$ *with* $\gamma > 0$, $\rho < 0$, *and auxiliary function* $A(t)$, *and* $h \in 2\mathrm{RV}_{\alpha,\beta}$ *with* $\alpha > 0$, $\beta < 0$ *and auxiliary function* $B(t)$. *Suppose that* $\int_1^\infty g(x^{-1/(\alpha\gamma+\delta)})\mathrm{d}x < \infty$ *for some* $\delta > 0$. *Then*

$$\rho_p[X] = \lambda_\gamma h\left(F^{\leftarrow}(p)\right)\left(1 + I_{\gamma,\rho} C\left(\frac{1}{1-p}\right)(1+o(1))\right),$$

where $\lambda_\gamma = 1 + \int_1^\infty g\left(x^{-1/(\alpha\gamma)}\right) \mathrm{d}x$, $C(t) = \alpha A(t) + \gamma B \circ U(t)$, $\eta = \max\{\rho, \beta\gamma\}$, *and*

$$I_{\gamma,\rho} = \begin{cases} -\frac{1}{\lambda_\gamma \eta}\left(\int_0^1 g\left(x^{-1/(\alpha\gamma+\eta)}\right) \mathrm{d}x + \int_1^\infty g\left(x^{-1/(\alpha\gamma)}\right) \mathrm{d}x\right), & \alpha\gamma \le |\eta|, \\ \frac{1}{\lambda_\gamma \eta}\int_1^\infty \left(g\left(x^{-1/(\alpha\gamma+\eta)}\right) - g\left(x^{-1/(\alpha\gamma)}\right)\right) \mathrm{d}x, & \alpha\gamma > |\eta|. \end{cases}$$

2.3 Estimations

In this section, we discuss the estimation of the generalized tail distortion risk measure based on the first-order asymptotic expansion in Theorem 2.2.1. That is we propose to estimate $\rho_{p_n}(X)$ by

$$\widehat{\rho}_{p_n}(X) = \widehat{\lambda}_\gamma h \circ \widehat{F}^{\leftarrow}(p_n), \tag{2.3.1}$$

where $\widehat{F}^{\leftarrow}$ is the estimator for the quantile function and $\widehat{\lambda}_\gamma = \left(1 + \int_1^\infty g\left(x^{-1/(\alpha\widehat{\gamma})}\right) \mathrm{d}x\right)$ for $\widehat{\gamma}$ being the estimator for the tail index γ. p_n can be assumed to be an intermediate sequence or an extreme sequence. We investigate the asymptotic normality of the estimator $\widehat{\rho}_{p_n}(X)$ for each case separately.

Note that El Methni and Stupfler (2017) also proposed an estimator based on the asymptotic expansions but under the assumption that $h(x) = x^\alpha$. In this paper, we give a complete treatment of the estimator in Eq. (2.3.1) for general h.

2.3.1 Intermediate level

In this subsection, we study the estimator in Eq. (2.3.1) at the intermediate level. Let q_n be an intermediate level sequence, that is a sequence satisfying that as $n \to \infty$, $q_n \to 1$ and $n(1 - q_n) \to \infty$. We study the property of $\widehat{\rho}_{q_n}(X)$.

At the intermediate level, the quantile function is estimated as

$$\widehat{F}^{\leftarrow}(q_n) = X_{n-\lfloor n(1-q_n)\rfloor,n}.$$

The tail index γ can be estimated by the Hill estimator, the moment estimator and other commonly used estimator for γ. Then we have the intermediate level estimator for the generalized tail distortion risk measure as

$$\widehat{\rho}_{q_n}(X) = \widehat{\lambda}_\gamma h\left(X_{n-\lfloor n(1-q_n)\rfloor,n}\right), \tag{2.3.2}$$

where

$$\widehat{\lambda}_\gamma = 1 + \int_1^\infty g\left(x^{-1/(\alpha\widehat{\gamma})}\right) \mathrm{d}x.$$

The next theorem shows the asymptotic normality of $\widehat{\rho}_{q_n}(X)$.

Theorem 2.3.1. *Under the conditions of Theorem 2.2.2, if* $q_n \uparrow 1, n(1 - q_n) \to \infty$, *and*

$\sqrt{n(1-q_n)}C\left(\frac{1}{1-q_n}\right) \to \zeta \in \mathbb{R}$, *and we assume further that*

$$\sqrt{n(1-q_n)}\left(\widehat{\gamma}-\gamma, \frac{\widehat{F}^{\leftarrow}(q_n)}{F^{\leftarrow}(q_n)}-1\right) \xrightarrow{d} (\Gamma, \Theta)$$

for some distributions (Γ, Θ)*, then as* $n \to \infty$*, we have*

$$\sqrt{n(1-q_n)}\left(\frac{\widehat{\rho}_{q_n}(X)}{\rho_{q_n}(X)}-1\right) \xrightarrow{d} m(\gamma)\Gamma + \alpha\Theta - \zeta I_{\gamma,\rho},$$

where

$$m(\gamma) = \frac{\int_1^\infty g'\left(x^{-1/(\alpha\gamma)}\right) x^{-1/(\alpha\gamma)} \ln x \, \mathrm{d}x}{\alpha\gamma^2\lambda_\gamma}.$$

Proof. By the definition of $\widehat{\rho}_{q_n}(X)$ and Theorem 2.2.2, we have

$$\begin{aligned}\frac{\widehat{\rho}_{q_n}(X)}{\rho_{q_n}(X)} &= \frac{\widehat{\lambda}_\gamma h\left(\widehat{F}^{\leftarrow}(q_n)\right)}{\lambda_\gamma h\left(F^{\leftarrow}(q_n)\right)\left(1+I_{\gamma,\rho}C\left(\frac{1}{1-q_n}\right)(1+o(1))\right)} \\ &= \frac{h\left(\widehat{F}^{\leftarrow}(q_n)\right)}{h\left(F^{\leftarrow}(q_n)\right)} \times \frac{1+\int_1^\infty g\left(x^{-1/(\alpha\widehat{\gamma})}\right)\mathrm{d}x}{1+\int_1^\infty g\left(x^{-1/(\alpha\gamma)}\right)\mathrm{d}x} \times \frac{1}{1+I_{\gamma,\rho}C\left(\frac{1}{1-q_n}\right)(1+o(1))}.\end{aligned}$$

By taking log on both sides, we obtain

$$\begin{aligned}\log\frac{\widehat{\rho}_{q_n}(X)}{\rho_{q_n}(X)} = \log\frac{h\left(\widehat{F}^{\leftarrow}(q_n)\right)}{h\left(F^{\leftarrow}(q_n)\right)} &+ \left(\log\left(1+\int_1^\infty g\left(x^{-1/(\alpha\widehat{\gamma})}\right)\mathrm{d}x\right) - \log\left(1+\int_1^\infty g\left(x^{-1/(\alpha\gamma)}\right)\mathrm{d}x\right)\right) \\ &- \log\left(1+I_{\gamma,\rho}C\left(\frac{1}{1-q_n}\right)(1+o(1))\right) := I_1 + I_2 - I_3.\end{aligned}$$

For I_1, noting that $h \in 2\mathrm{RV}_{\alpha,\beta}$, we rewrite the 2RV definition as

$$\lim_{t\to\infty}\frac{x^{-\alpha}\frac{h(tx)}{h(t)}-1}{B(t)} = \frac{x^\beta-1}{\beta}, \qquad x>0. \tag{2.3.3}$$

Since $\lim_{t\to\infty} B(t) = 0$, Eq. (2.3.3) is equivalent to

$$\lim_{t\to\infty}\frac{\log h(tx) - \log h(t) - \alpha\log x}{B(t)} = \frac{x^\beta-1}{\beta}.$$

By Theorem 2.3.9 of Haan and Ferreira (2006), there exists $B_0(t) \sim B(t)$ such that for any $\varepsilon > 0$, there exists $t_0 > 0$ such that for all $t \geq t_0$, $x \geq 1$,

$$\left| \frac{\log h(tx) - \log h(t) - \alpha \log x}{B_0(t)} - \frac{x^\beta - 1}{\beta} \right| \leq \varepsilon x^{\beta+\varepsilon}. \tag{2.3.4}$$

Since $\widehat{F}^{\leftarrow}(q_n) = X_{n-\lfloor n(1-q_n) \rfloor, n}$ as $n \to \infty$, and by letting $t = F^{\leftarrow}(q_n)$ and $x = \frac{\widehat{F}^{\leftarrow}(q_n)}{F^{\leftarrow}(q_n)}$,

$$\log \frac{h\left(\widehat{F}^{\leftarrow}(q_n)\right)}{h\left(F^{\leftarrow}(q_n)\right)} = \alpha \log \frac{\widehat{F}^{\leftarrow}(q_n)}{F^{\leftarrow}(q_n)} + B_0(F^{\leftarrow}(q_n)) \frac{\left(\frac{\widehat{F}^{\leftarrow}(q_n)}{F^{\leftarrow}(q_n)}\right)^\beta - 1}{\beta} + o_P(1) B_0(F^{\leftarrow}(q_n)) \left(\frac{\widehat{F}^{\leftarrow}(q_n)}{F^{\leftarrow}(q_n)}\right)^{\beta+\varepsilon}.$$

By the assumption that

$$\sqrt{n(1-q_n)} \left(\frac{\widehat{F}^{\leftarrow}(q_n)}{F^{\leftarrow}(q_n)} - 1 \right) \xrightarrow{d} \Theta,$$

we obtain as $n \to \infty$

$$\sqrt{n(1-q_n)} \log \frac{\widehat{F}^{\leftarrow}(q_n)}{F^{\leftarrow}(q_n)} \xrightarrow{d} \Theta,$$

and

$$\frac{\left(\frac{\widehat{F}^{\leftarrow}(q_n)}{F^{\leftarrow}(q_n)}\right)^\beta - 1}{\beta} \xrightarrow{P} 0, \qquad \left(\frac{\widehat{F}^{\leftarrow}(q_n)}{F^{\leftarrow}(q_n)}\right)^{\beta+\varepsilon} \xrightarrow{P} 1.$$

Lastly, since $\sqrt{n(1-q_n)} C\left(\frac{1}{1-q_n}\right) \to \zeta$, it follows that $\sqrt{n(1-q_n)} B_0(F^{\leftarrow}(q_n)) \to \widetilde{\zeta} \in \mathbb{R}$. Thus,

$$\sqrt{n(1-q_n)} I_1 \xrightarrow{d} \alpha\Theta.$$

For I_2, note that

$$\frac{d}{dt} \log \left(1 + \int_1^\infty g\left(x^{-1/(\alpha t)}\right) \mathrm{d}x \right) = \frac{\int_1^\infty g'\left(x^{-1/(\alpha\gamma)}\right) x^{-1/(\alpha\gamma)} \ln x \, \mathrm{d}x}{\alpha\gamma^2 \lambda_\gamma}.$$

By the delta-method as $n \to \infty$, it follows that

$$\sqrt{n(1-q_n)} I_2 \xrightarrow{d} \frac{\int_1^\infty g'\left(x^{-1/(\alpha\gamma)}\right) x^{-1/(\alpha\gamma)} \ln x \, \mathrm{d}x}{\alpha\gamma^2 \lambda_\gamma} \Gamma.$$

For I_3, note that as $n \to \infty$, $C\left(\frac{1}{1-q_n}\right) \to 0$ and

$$I_3 = I_{\gamma,\rho} C\left(\frac{1}{1-q_n}\right)(1+o(1)).$$

Then,

$$\sqrt{n(1-q_n)} I_3 \xrightarrow{d} \zeta I_{\gamma,\rho}.$$

Combining all the terms above yields

$$\sqrt{n(1-q_n)}\log\frac{\widehat{\rho}_{q_n}(X)}{\rho_{q_n}(X)} \xrightarrow{d} \alpha\Theta + \frac{\int_1^\infty g'\left(x^{-1/(\alpha\gamma)}\right)x^{-1/(\alpha\gamma)}\ln x\,\mathrm{d}x}{\alpha\gamma^2\lambda_\gamma}\Gamma - \zeta I_{\gamma,\rho}. \qquad \square$$

In particular, when γ is estimated by the Hill estimator with $k = \lfloor n(1-q_n)\rfloor$, i.e.

$$\widehat{\gamma}_H = \frac{1}{k}\sum_{i=0}^{k-1}\log X_{n-i,n} - \log X_{n-k,n},$$

we denote the estimator of the generalized tail distortion risk measure by $\widehat{\rho}^H_{q_n}(X)$. Then by Theorems 2.4.1 and 3.2.5, and Lemma 3.2.3 of Haan and Ferreira (2006),

$$\sqrt{n(1-q_n)}\left(\widehat{\gamma}_H - \gamma, \frac{X_{n-\lfloor n(1-q_n)\rfloor,n}}{F^{\leftarrow}(q_n)} - 1\right) \xrightarrow{d} (\Gamma,\Theta),$$

where $\Gamma = N\left(\frac{\zeta}{1-\rho},\gamma^2\right)$ is independent of $\Theta = N(0,\gamma^2)$. We have the following simpler asymptotic normality of $\widehat{\rho}_{q_n}(X)$.

Corollary 2.3.1. *Under the conditions of Theorem 2.2.2, if $q_n \uparrow 1, n(1-q_n) \to \infty$, and $\sqrt{n(1-q_n)}C\left(\frac{1}{1-q_n}\right) \to \zeta \in \mathbb{R}$, then we have the following as $n \to \infty$:*

$$\sqrt{n(1-q_n)}\left(\frac{\widehat{\rho}^H_{q_n}(X)}{\rho_{q_n}(X)} - 1\right) \xrightarrow{d} N\left(\frac{m(\gamma)\zeta}{1-\rho} - \zeta I_{\gamma,\rho}, \left(\left(m(\gamma)\right)^2 + \alpha^2\right)\gamma^2\right).$$

2.3.2 Extreme level

In this subsection, we study the estimator in Eq. (2.3.1) at the extreme level. Let τ_n be an extreme level sequence, that is a sequence satisfying $n \to \infty$, $\tau_n \uparrow 1$ and $n(1-\tau_n) \to c$ for some $c < \infty$. We study the property of $\widehat{\rho}_{\tau_n}(X)$.

Let $\{\tau_n, n \in N\}$ and $\{q_n, n \in N\}$ be two sequences of levels satisfying $\tau_n, q_n \uparrow 1, n(1-q_n) \to \infty$, and $n(1-\tau_n) \to c < \infty$. Then, by the assumption that $U \in \mathrm{RV}_\gamma$ and $h \in \mathrm{RV}_\alpha$, we have $h \circ F^{\leftarrow}\left(1 - \frac{1}{\cdot}\right) \in \mathrm{RV}_{\alpha\gamma}$ and hence

$$\frac{\rho_{\tau_n}(X)}{\rho_{q_n}(X)} \sim \frac{h \circ F^{\leftarrow}(\tau_n)}{h \circ F^{\leftarrow}(q_n)} \sim \left(\frac{1-\tau_n}{1-q_n}\right)^{-\alpha\gamma}, \qquad n \to \infty.$$

Thus we propose to estimate $\rho_{\tau_n}(X)$ as

$$\widehat{\rho}^*_{\tau_n}(X) = \left(\frac{1-\tau_n}{1-q_n}\right)^{-\alpha\widehat{\gamma}} \widehat{\rho}_{q_n}(X). \tag{2.3.5}$$

Theorem 2.3.2. *Under the conditions of Theorem 2.2.2, if $\tau_n, q_n \uparrow 1, n(1-q_n) \to \infty$, and $n(1-\tau_n) \to c < \infty$ and we assume further that*

$$\sqrt{n(1-q_n)}\left(\frac{\widehat{\rho}_{q_n}(X)}{\rho_{q_n}(X)} - 1\right) \xrightarrow{d} \Delta \quad \textit{and} \quad \sqrt{n(1-q_n)}(\widehat{\gamma} - \gamma) \xrightarrow{d} \Gamma$$

with $\sqrt{n(1-q_n)}C((1-q_n)^{-1}) \to \zeta \in \mathbb{R}$ and $\sqrt{n(1-q_n)}/\log[(1-q_n)/(1-\tau_n)] \to \infty$, then

$$\frac{\sqrt{n(1-q_n)}}{\log[(1-q_n)/(1-\tau_n)]}\left(\frac{\widehat{\rho}^*_{\tau_n}(X)}{\rho_{\tau_n}(X)} - 1\right) \xrightarrow{d} \alpha\Gamma.$$

Proof. By the definition of $\widehat{\rho}^*_{\tau_n}(X)$, it follows that

$$\log\frac{\widehat{\rho}^*_{\tau_n}(X)}{\rho_{\tau_n}(X)} = \log\left(\frac{\left(\frac{1-\tau_n}{1-q_n}\right)^{-\alpha\widehat{\gamma}}\widehat{\rho}_{q_n}(X)}{\rho_{\tau_n}(X)}\right)$$

$$= \log\left(\frac{\left(\frac{1-\tau_n}{1-q_n}\right)^{-\alpha\widehat{\gamma}} h\left(\widehat{F}^{\leftarrow}(q_n)\right)}{h\left(F^{\leftarrow}(\tau_n)\right)}\right) + \log\frac{\widehat{\rho}_{q_n}(X)}{\rho_{q_n}(X)} + \log\frac{\rho_{q_n}(X)}{h\left(F^{\leftarrow}(q_n)\right)}$$
$$- \log\frac{h\left(\widehat{F}^{\leftarrow}(q_n)\right)}{h\left(F^{\leftarrow}(q_n)\right)} - \log\frac{\rho_{\tau_n}(X)}{h\left(F^{\leftarrow}(\tau_n)\right)}. \tag{2.3.6}$$

We only need to analyze the first term in Eq. (2.3.6) as the other terms follow the similar arguments in the proof of Theorem 3.4 of Zhao, Mao, and Yang (2021).

First note the rewriting,

$$\frac{\left(\frac{1-\tau_n}{1-q_n}\right)^{-\alpha\widehat{\gamma}} h\left(\widehat{F}^{\leftarrow}(q_n)\right)}{h\left(F^{\leftarrow}(\tau_n)\right)} = \left(\frac{1-\tau_n}{1-q_n}\right)^{-\alpha\widehat{\gamma}} \frac{h\left(\widehat{F}^{\leftarrow}(q_n)\right)}{h\left(\left(\frac{1-\tau_n}{1-q_n}\right)^{-\widehat{\gamma}} \widehat{F}^{\leftarrow}(q_n)\right)} \frac{h\left(\left(\frac{1-\tau_n}{1-q_n}\right)^{-\widehat{\gamma}} \widehat{F}^{\leftarrow}(q_n)\right)}{h\left(F^{\leftarrow}(\tau_n)\right)}.$$

Since $h \in 2\mathrm{RV}_{\alpha,\beta}$, by Eq. (2.3.4), we have

$$\begin{aligned}
\log\left(\frac{\left(\frac{1-\tau_n}{1-q_n}\right)^{-\alpha\widehat{\gamma}} h\left(\widehat{F}^{\leftarrow}(q_n)\right)}{h\left(F^{\leftarrow}(\tau_n)\right)}\right) &= \alpha\log\left(\frac{\left(\frac{1-\tau_n}{1-q_n}\right)^{-\widehat{\gamma}} \widehat{F}^{\leftarrow}(q_n)}{F^{\leftarrow}(\tau_n)}\right) + B_0(F^{\leftarrow}(\tau_n))\frac{\left(\frac{\left(\frac{1-\tau_n}{1-q_n}\right)^{-\widehat{\gamma}} \widehat{F}^{\leftarrow}(q_n)}{F^{\leftarrow}(\tau_n)}\right)^{\beta} - 1}{\beta} \\
&\quad + o_p(1)B_0(F^{\leftarrow}(\tau_n))\left(\frac{\left(\frac{1-\tau_n}{1-q_n}\right)^{-\widehat{\gamma}} \widehat{F}^{\leftarrow}(q_n)}{F^{\leftarrow}(\tau_n)}\right)^{(\beta+\varepsilon)} \\
&\quad - B_0(\widehat{F}^{\leftarrow}(q_n))\frac{\left(\frac{1-\tau_n}{1-q_n}\right)^{-\widehat{\gamma}\beta} - 1}{\beta} + o_p(1)B_0(\widehat{F}^{\leftarrow}(q_n))\left(\frac{1-\tau_n}{1-q_n}\right)^{-\widehat{\gamma}(\beta+\varepsilon)}.
\end{aligned}$$

By Theorem 4.3.8 of Haan and Ferreira (2006), we obtain

$$\frac{\sqrt{n(1-q_n)}}{\log[(1-q_n)/(1-\tau_n)]}\log\left(\frac{\left(\frac{1-\tau_n}{1-q_n}\right)^{-\widehat{\gamma}} \widehat{F}^{\leftarrow}(q_n)}{F^{\leftarrow}(\tau_n)}\right) \xrightarrow{d} \Gamma,$$

and

$$\frac{\left(\frac{\left(\frac{1-\tau_n}{1-q_n}\right)^{-\widehat{\gamma}}\widehat{F}^{\leftarrow}(q_n)}{F^{\leftarrow}(\tau_n)}\right)^{\beta}-1}{\beta}\xrightarrow{P}0, \qquad \left(\frac{\left(\frac{1-\tau_n}{1-q_n}\right)^{-\widehat{\gamma}}\widehat{F}^{\leftarrow}(q_n)}{F^{\leftarrow}(\tau_n)}\right)^{(\beta+\varepsilon)}\xrightarrow{P}1.$$

Note that $\frac{1-\tau_n}{1-q_n}\to 0$. Then by $B_0\in \mathrm{RV}_\beta$ with $\beta<0$ and $U\in \mathrm{RV}_\gamma$, it follows that as $n\to\infty$,

$$\frac{B_0(F^{\leftarrow}(\tau_n))}{B_0(F^{\leftarrow}(q_n))}\sim\left(\frac{1-\tau_n}{1-q_n}\right)^{-\gamma\beta}\to 0,$$

and

$$\frac{B_0(\widehat{F}^{\leftarrow}(q_n))}{B_0(F^{\leftarrow}(q_n))}\xrightarrow{P}1.$$

Finally, by noting that $\sqrt{n(1-q_n)}B_0(F^{\leftarrow}(q_n))\to\widetilde{\zeta}$, which yields $\frac{\sqrt{n(1-q_n)}}{\log[(1-q_n)/(1-\tau_n)]}B_0(F^{\leftarrow}(q_n))\to 0$, we have

$$\frac{\sqrt{n(1-q_n)}}{\log[(1-q_n)/(1-\tau_n)]}\log\left(\frac{\left(\frac{1-\tau_n}{1-q_n}\right)^{-\alpha\widehat{\gamma}}h\left(\widehat{F}^{\leftarrow}(q_n)\right)}{h\left(F^{\leftarrow}(\tau_n)\right)}\right)\xrightarrow{d}\alpha\Gamma.$$

□

Similarly, when γ is estimated by the Hill estimator $\widehat{\gamma}_H$, we denote the extreme level estimator of the generalized tail distortion risk measure as $\widehat{\rho}^{*H}_{\tau_n}(X)$. Then we have the following simpler asymptotic normality of $\widehat{\rho}^{*H}_{\tau_n}(X)$.

Corollary 2.3.2. *Under the conditions of Theorem 2.2.2, if $\tau_n, q_n\uparrow 1, n(1-q_n)\to\infty$, and $n(1-\tau_n)\to c<\infty$, and we further assume that*

$$\sqrt{n(1-q_n)}\left(\frac{\widehat{\rho}_{q_n}(X)}{\rho_{q_n}(X)}-1\right)\xrightarrow{d}\Delta$$

with $\sqrt{n(1-q_n)}C((1-q_n)^{-1})\to\zeta\in\mathbb{R}$ and $\sqrt{n(1-q_n)}/\log[(1-q_n)/(1-\tau_n)]\to\infty$, then we have the following as $n\to\infty$

$$\frac{\sqrt{n(1-q_n)}}{\log[(1-q_n)/(1-\tau_n)]}\left(\frac{\widehat{\rho}^{*H}_{\tau_n}(X)}{\rho_{q_n}(X)}-1\right)\xrightarrow{d}N\left(\frac{\alpha\zeta}{1-\rho},\alpha^2\gamma^2\right).$$

2.3.3 Bias-corrected estimations

In this subsection, we consider bias-corrected estimation at both the intermediate and extreme levels. Here we focus on the case that the tail index γ is estimated by the Hill estimator.

Due to Corollary 2.3.1, we propose the bias-corrected estimator at the intermediate level as

$$\widehat{\rho}_{q_n}^{HB}(X) = \frac{\widehat{\rho}_{q_n}^{H}(X)}{1+\widehat{\varepsilon}^{H}},$$

where

$$\widehat{\varepsilon}^{H} = \left(\frac{m(\widehat{\gamma}_H)}{1-\widehat{\rho}} - I_{\widehat{\gamma}_H,\widehat{\rho}}\right)\widehat{C}\left(\frac{1}{1-q_n}\right).$$

Since $C(t) = \alpha A(t) + \gamma B \circ U(t)$, we can estimate $C(t)$ as

$$\widehat{C}\left(\frac{1}{1-q_n}\right) = \alpha\widehat{A}\left(\frac{1}{1-q_n}\right) + \widehat{\gamma}_H\widehat{B}\left(\widehat{F}^{\leftarrow}(q_n)\right).$$

Suppose $\widehat{A}$, $\widehat{B}$ and $\widehat{\rho}$ are consistent estimators of A, B and ρ, for example, A and B can be estimated using the formulation in Haouas, Necir, and Brahimi (2019) and ρ can be estimated using Eq. (12) in Troop, Godin, and Yu (2021). To be more specific, we estimate ρ as

$$\widehat{\rho} = \min\left\{0, \frac{3\left(T_{n,k}^{(\tau)}-1\right)}{T_{n,k}^{(\tau)}-3}\right\},$$

where

$$T_{n,k}^{(\tau)} = \frac{\left(M_{n,k}^{(1)}\right)^{\tau} - \left(M_{n,k}^{(2)}/2\right)^{\tau/2}}{\left(M_{n,k}^{(2)}/2\right)^{\tau/2} - \left(M_{n,k}^{(3)}/6\right)^{\tau/3}}, \quad \tau \in \mathbb{R},$$

and

$$M_{n,k}^{(j)} = \frac{1}{k}\sum_{i=0}^{k-1}\left(\log X_{n-i,n} - \log X_{n-k,n}\right)^{j}, \quad j > 0.$$

The auxiliary function $A(t)$ is estimated as

$$\widehat{A}(t) = \frac{(1-\widehat{\rho})^2 \left(M_{n,k}^{(2)} - 2\left(M_{n,k}^{(1)}\right)^2 \right)}{2\widehat{\rho} M_{n,k}^{(1)}}.$$

The next theorem shows that $\widehat{\rho}_{q_n}^{HB}(X)$ is asymptotically unbiased.

Theorem 2.3.3. *Under the conditions of Corollary 2.3.1, we further assume that as $n \to \infty$,*

$$\frac{\widehat{A}}{A(n/k)} \xrightarrow{P} 1, \qquad \frac{\widehat{B}}{B(n/k)} \xrightarrow{P} 1, \qquad \widehat{\rho} \xrightarrow{P} \rho.$$

Then we have as $n \to \infty$,

$$\sqrt{n(1-q_n)} \left(\frac{\widehat{\rho}_{q_n}^{HB}(X)}{\rho_{q_n}(X)} - 1 \right) \xrightarrow{d} N\left(0, \left((m(\gamma))^2 + \alpha^2\right)\gamma^2\right).$$

Proof. Note that

$$\sqrt{n(1-q_n)} \left(\frac{\widehat{\rho}_{q_n}^{HB}(X)}{\rho_{q_n}(X)} - 1 \right) = \sqrt{n(1-q_n)} \left(\frac{\widehat{\rho}_{q_n}^{H}(X)}{\rho_{q_n}(X)} - 1 \right) - \sqrt{n(1-q_n)}\widehat{\varepsilon}^H \frac{\widehat{\rho}_{q_n}^{H}(X)}{\rho_{q_n}(X)}.$$

Since $\widehat{\rho}_{q_n}^{H}(X)$, $\widehat{\gamma}_H$, $\widehat{A}$, $\widehat{B}$ and $\widehat{\rho}$ are consistent and $\sqrt{n(1-q_n)}C\left(\frac{1}{1-q_n}\right) \to \zeta$, we have

$$\sqrt{n(1-q_n)}\widehat{\varepsilon}^H \frac{\widehat{\rho}_{q_n}^{H}(X)}{\rho_{q_n}(X)} \xrightarrow{P} \frac{m(\gamma)\zeta}{1-\rho} - \zeta I_{\gamma,\rho}.$$

Thus, $\widehat{\rho}_{q_n}^{HB}(X)$ is asymptotically unbiased. □

At the extreme level, due to Corollary 2.3.2, we propose the following bias-corrected estimator

$$\widehat{\rho}_{\tau_n}^{*HB}(X) = \frac{\widehat{\rho}_{\tau_n}^{*H}(X)}{(1+\widehat{\varepsilon}^{*H})},$$

where

$$\widehat{\varepsilon}^{*H} = \frac{\alpha \widehat{C}\left(\frac{1}{1-q_n}\right)}{1-\widehat{\rho}} \log\left(\frac{1-q_n}{1-\tau_n}\right).$$

Similarly we can show that $\widehat{\rho}^{*HB}_{\tau_n}(X)$ is asymptotically unbiased.

Theorem 2.3.4. *Under the conditions of Corollary 2.3.1, we further assume that as* $n \to \infty$,

$$\frac{\widehat{A}}{A(n/k)} \xrightarrow{P} 1, \qquad \frac{\widehat{B}}{B(n/k)} \xrightarrow{P} 1, \qquad \widehat{\rho} \xrightarrow{P} \rho.$$

Then we have as $n \to \infty$,

$$\frac{\sqrt{n(1-q_n)}}{\log[(1-q_n)/(1-\tau_n)]}\left(\frac{\widehat{\rho}^{*HB}_{\tau_n}(X)}{\rho_{q_n}(X)} - 1\right) \xrightarrow{d} N\left(0, \alpha^2\gamma^2\right).$$

2.4 Simulations

In this section, we explore the performances of the estimators proposed in the last section. We compare our estimators with the ones in the literature. That is we compare $\widehat{\rho}^{H}_{q_n}(X)$ with $\widehat{R}^{PL}_{q_n}$ and $\widehat{R}^{Wins}_{q_n}(\psi_n)$, in (El Methni and Stupfler 2017, 2018), respectively. We also compare $\widehat{\rho}^{*H}_{\tau_n}(X)$ with $\widehat{R}^{PL}_{\tau_n}$ and $\widehat{R}^{Wins}_{\tau_n}(\psi_n)$, in El Methni and Stupfler (2017, 2018), respectively. The empirical quantile functions give rise to PL estimator $(\widehat{R}^{PL}_{q_n})$ if $\psi_n(s) = 1-(1-q_n)s$ and Wins-PL estimator $(\widehat{R}^{Wins}_{q_n})$ if $\psi_n(s) = \min(t_n, 1-(1-q_n)s)$. Thus, the estimators we compare against are

$$\begin{aligned}
\widehat{R}^{PL}_{q_n}\left(h(X)\right) &:= \int_0^1 h \circ \widehat{F}^{\leftarrow}(1-(1-q_n)s)\mathrm{d}g(s),\\
\widehat{R}^{Wins}_{q_n}\left(h(X)\right) &:= \int_0^1 h \circ \widehat{F}^{\leftarrow}(\min(t_n, 1-(1-q_n)s))\mathrm{d}g(s),\\
\widehat{R}^{PL}_{\tau_n}\left(h(X)\right) &:= \left(\frac{1-\tau_n}{1-q_n}\right)^{-\alpha\widehat{\gamma}} \widehat{R}^{PL}_{q_n}\left(h(X)\right),\\
\widehat{R}^{Wins}_{\tau_n}\left(h(X)\right) &:= \left(\frac{1-\tau_n}{1-q_n}\right)^{-\alpha\widehat{\gamma}} \widehat{R}^{Wins}_{q_n}\left(h(X)\right),
\end{aligned}$$

where the $t_n = 0.995$

Consider Pareto$(1/\gamma)$ distribution with distribution function given by

$$F(x) = 1 - \left(\frac{1}{x+1}\right)^{1/\gamma}, \qquad x > 0,$$

for some $\gamma > 0$. One can verify that the tail quantile function of the Pareto$(1/\gamma)$ distribution satisfies that $U(t) \in 2\mathrm{RV}_{\gamma,-\gamma}$ with auxiliary function $A(t) = \gamma t^{-\gamma}$.

In order to compare with the estimators in El Methni and Stupfler (2017, 2018), let $h(t) = t^\alpha$ for some $\alpha > 0$. Thus, $h \in 2\mathrm{RV}_{\alpha,-\infty}$ with an auxiliary function $B(t) = 0$. In addition, let $g(x) = x$.

2.4.1 Intermediate level

We simulate 1000 replications of random samples from the Pareto$(1/\gamma)$ distribution. All the experiments use sample sizes of $n = 1000, 5000, 20000$ and 50000 and the true index of the Pareto distribution $\gamma = 1/3, 1/5$. Moreover, we pick the $\alpha = 1.2$. The choices of γ and α satisfy $\alpha < 1/\gamma$. It is well known that choosing the value of k is always challenging in the field of extreme value statistics and there is a bias-variance trade-off. Hence, we use a common method in extreme value statistics by choose k from the first stable region of Hill estimator. By the Hill plot, we take $q_n = 1 - k/n$ with k values that correspond to $q_n = 0.96$ and 0.97. In Table 2.1, we report the mean, sample standard deviation and relative mean squared errors (MSEs) of the ratio of the estimator to its true value, where we conclude that he lighter the tail of the distribution, the better the estimators. In Figure 2.1, we plot the MSEs and the sample standard deviations against different q_n levels, for samples of size 20000. We also explore the effect of α in the distortion function $h(t) = t^\alpha$. In Figure 2.2, we plot them against difference values of α, for samples of size 20000, respectively. Our estimator outperforms the Wins-PL estimator and approaches the accuracy of the PL

estimator for larger values of q_n and smaller values of α.

$(n, 1/\gamma, k)$	q_n	ρ_{q_n}	$\frac{\widehat{\rho}^H_{q_n}}{\rho_{q_n}}$	$\frac{\widehat{R}^{PL}_{q_n}}{\rho_{q_n}}$	$\frac{\widehat{R}^{Wins}_{q_n}}{\rho_{q_n}}$	MSE(H)	MSE(PL)	MSE(Wins)
(1000, 3, 30)	0.9700	5.1921	1.0972 (0.2618)	1.0043 (0.2241)	0.8279 (0.1148)	0.0779	0.0502	0.0428
(1000, 3, 40)	0.9600	4.4899	1.1006 (0.2184)	1.0048 (0.1998)	0.8518 (0.1084)	0.0577	0.0399	0.0337
(1000, 5, 30)	0.9700	1.6819	1.0482 (0.1453)	0.9997 (0.1265)	0.8949 (0.0942)	0.0234	0.0160	0.0199
(1000, 5, 40)	0.9600	1.4985	1.0547 (0.1301)	1.0006 (0.1138)	0.9123 (0.0884)	0.0199	0.0129	0.0155
(5000, 3, 150)	0.9700	5.1921	1.0670 (0.1000)	1.0074 (0.1135)	0.8288 (0.0520)	0.0145	0.0129	0.0320
(5000, 3, 200)	0.9600	4.4899	1.0760 (0.0915)	1.0068 (0.1007)	0.8519 (0.0491)	0.0141	0.0102	0.0244
(5000, 5, 150)	0.9700	1.6819	1.0411 (0.0629)	1.0035 (0.0566)	0.8972 (0.0430)	0.0056	0.0032	0.0124
(5000, 5, 200)	0.9600	1.4985	1.0475 (0.0579)	1.0033 (0.0511)	0.9138 (0.0403)	0.0056	0.0026	0.0091
(20000, 3, 600)	0.9700	5.1921	1.0547 (0.0470)	0.9989 (0.0483)	0.8254 (0.0251)	0.0052	0.0023	0.0311
(20000, 3, 800)	0.9600	4.4899	1.0665(0.0439)	0.9991 (0.0430)	0.8486 (0.0237)	0.0064	0.0019	0.0235
(20000, 5, 600)	0.9700	1.6819	1.0351 (0.0300)	0.9992 (0.0272)	0.8948 (0.0208)	0.0021	0.0007	0.0115
(20000, 5, 800)	0.9600	1.4985	1.0428 (0.0281)	0.9994 (0.0246)	0.9115 (0.0195)	0.0026	0.0006	0.0082
(50000, 3, 1500)	0.9700	5.1921	1.0550 (0.0311)	1.0007 (0.0317)	0.8263 (0.0164)	0.0040	0.0010	0.0305
(50000, 3, 2000)	0.9600	4.4899	1.0674 (0.0285)	1.0006 (0.0282)	0.8493 (0.0155)	0.0054	0.0008	0.0229
(50000, 5, 1500)	0.9700	1.6819	1.0356 (0.0198)	1.0003 (0.0178)	0.8956 (0.0136)	0.0017	0.0003	0.0111
(50000, 5, 2000)	0.9600	1.4985	1.0435 (0.0182)	1.0002 (0.0160)	0.9121 (0.0127)	0.0022	0.0003	0.0079

Table 2.1: Based on the sample of size $n = 1000, 5000, 20000$ and 50000 from the Pareto$(1/\gamma)$ distribution, the true values of the extreme Wang distortion risk measure ρ_{q_n} and the sample means, sample standard deviations and MSEs of $\frac{\widehat{\rho}^H_{q_n}}{\rho_{q_n}}, \frac{\widehat{R}^{PL}_{q_n}}{\rho_{q_n}}, \frac{\widehat{R}^{Wins}_{q_n}}{\rho_{q_n}}$ are presented for various values of k.

2.4.2 Extreme level

We simulate 1000 replications of random samples from the Pareto$(1/\gamma)$ distribution. All the experiments use sample sizes of $n = 1000, 5000, 20000$ and 50000 and the true index of the Pareto distribution $\gamma = 1/3, 1/5$. Moreover, we pick the $\alpha = 1.2$, the extreme level $\tau = 0.9995$ and the same q_n satisfy $\sqrt{n(1-q_n)}C((1-q_n)^{-1}) \to \zeta \in \mathbb{R}$. In Table 2.2, we report the mean, sample standard deviation and MSEs of the ratio of the estimator to its true value. In Figure 2.3, we plot the MSEs and the sample standard deviations against different q_n levels for samples of size 20000, respectively. The Wins-PL is the best performing estimator for the extreme level, our estimator is a close contender to the PL estimator

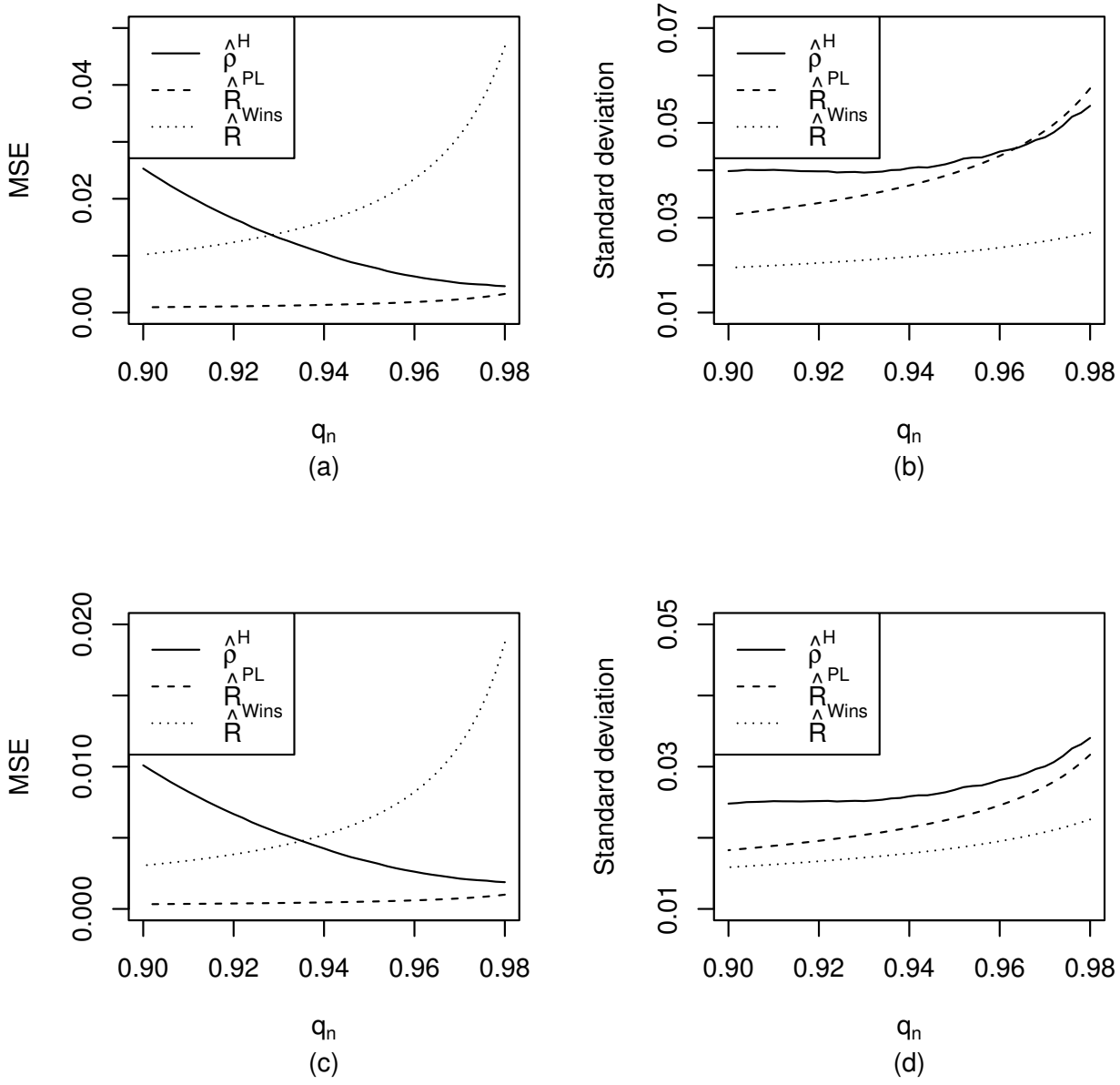


Figure 2.1: The MSEs (a,c) and sample standard deviations (b,d) of $\frac{\widehat{\rho}^{H}_{q_n}}{\rho_{q_n}}$ (solid line) , $\frac{\widehat{R}^{PL}_{q_n}}{\rho_{q_n}}$ (dashed line), $\frac{\widehat{R}^{Wins}_{q_n}}{\rho_{q_n}}$ (dash-dotted line) are plotted against q_n based on 1000 repetitions sampling 20000 times from the Pareto(3) and Pareto(5). The (a) and (b) are MSE and sample standard deviations of Pareto (3) and the (c) and (d) are MSE and sample standard deviations of Pareto(5).

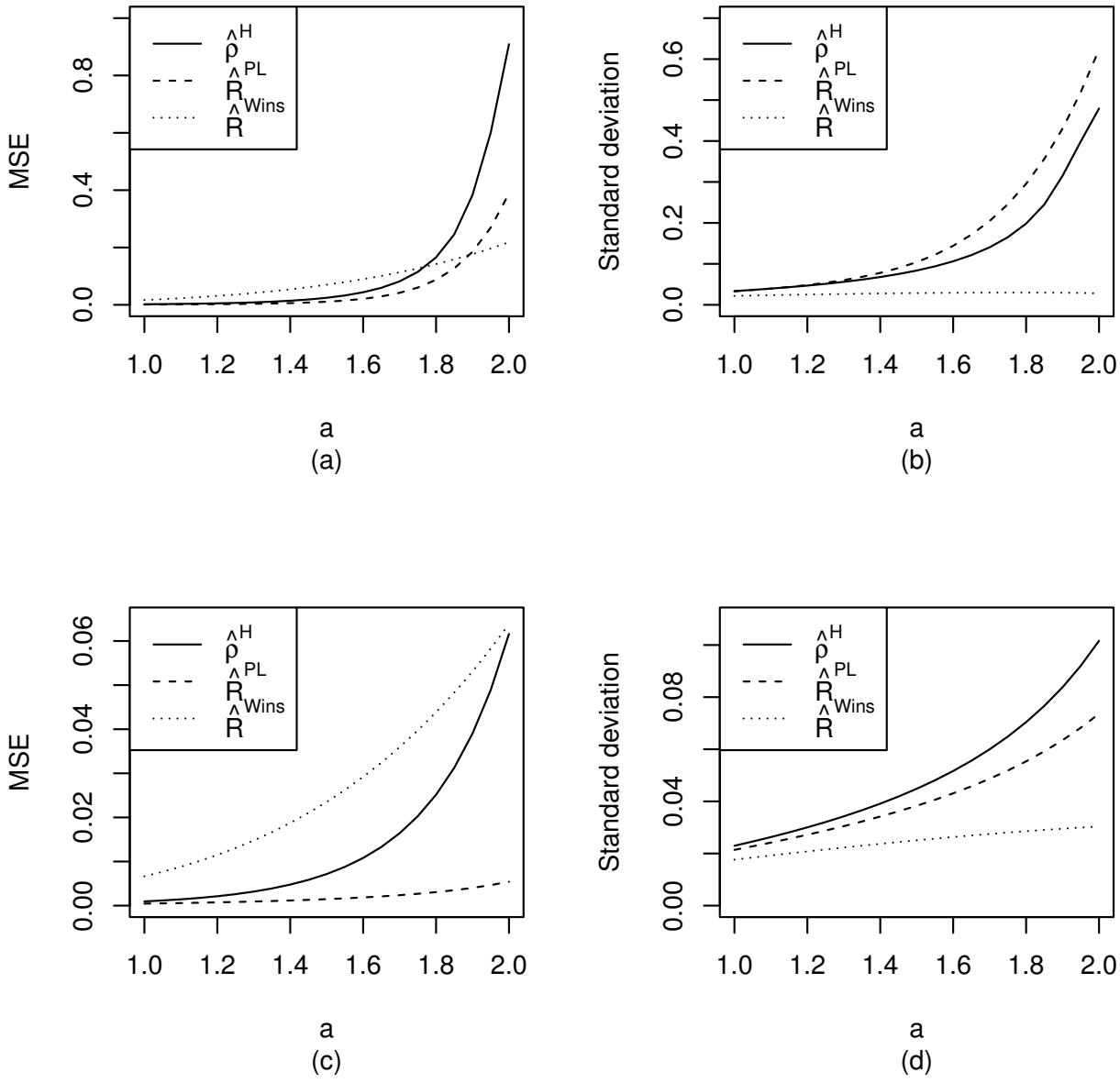


Figure 2.2: The MSEs (a,c) and sample standard deviations (b,d) of $\frac{\widehat{\rho}^{H}_{q_n}}{\rho_{q_n}}$ (solid line) , $\frac{\widehat{R}^{PL}_{q_n}}{\rho_{q_n}}$ (dashed line), $\frac{\widehat{R}^{Wins}_{q_n}}{\rho_{q_n}}$ (dash-dotted line) are plotted against α based on 1000 repetitions sampling 20000 times from the Pareto(3) and Pareto(5). The (a) and (b) are MSE and sample standard deviations of Pareto(3) and the (c) and (d) are MSE and sample standard deviations of Pareto(5).

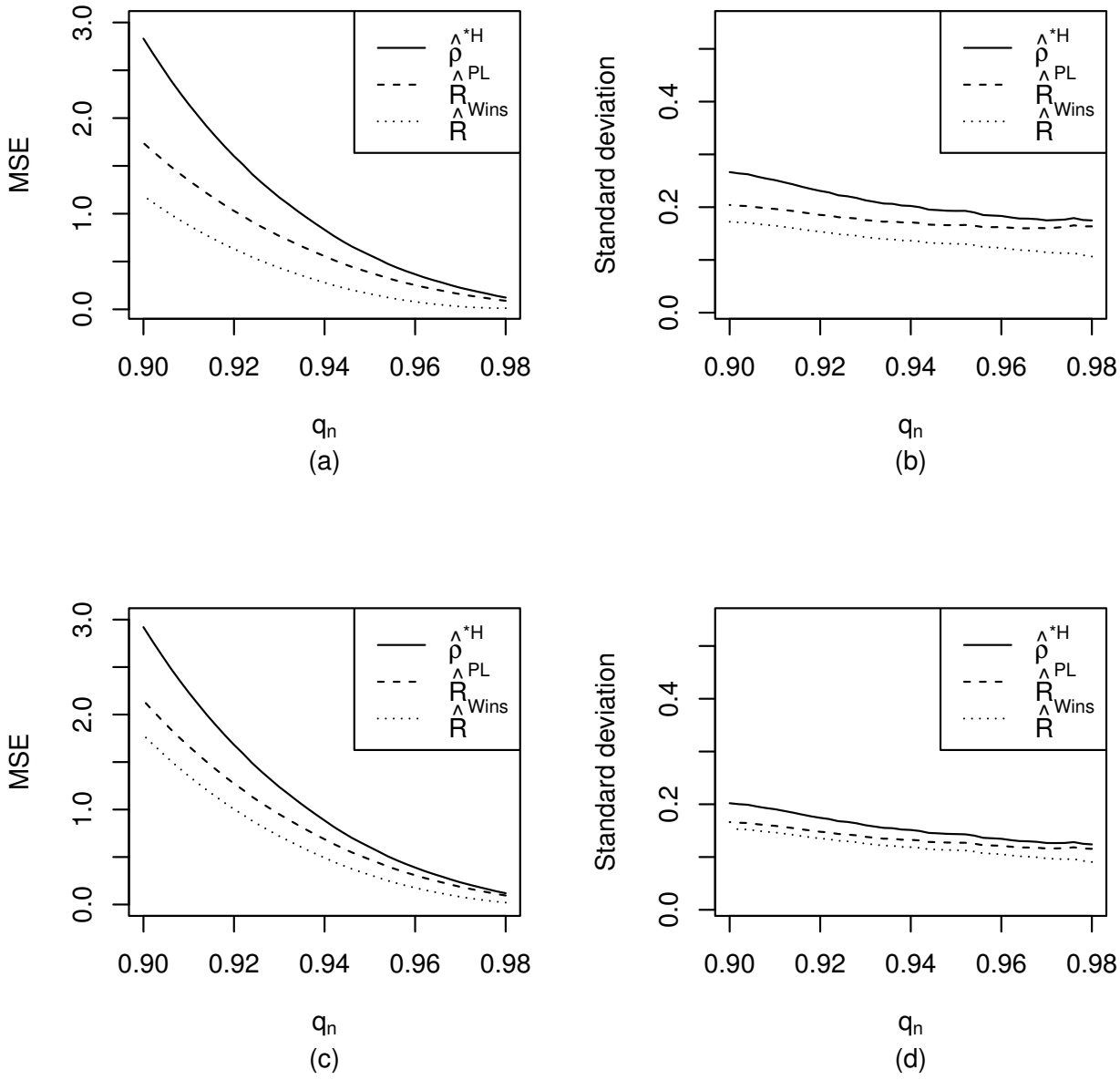


Figure 2.3: The MSEs (a,c) and sample standard deviations (b,d) of $\frac{\widehat{\rho}^{*H}_{\tau_n}}{\rho_{\tau_n}}$ (solid line) , $\frac{\widehat{R}^{PL}_{\tau_n}}{\rho_{\tau_n}}$ (dashed line), $\frac{\widehat{R}^{Wins}_{\tau_n}}{\rho_{\tau_n}}$ (dash-dotted line) are plotted against q_n based on 1000 repetitions sampling 20000 times from the Pareto(3) and Pareto(5). The (a) and (b) are MSE and sample standard deviations of Pareto (3) and the (c) and (d) are MSE and sample standard deviations of Pareto(5).

$(n, 1/\gamma, k)$	q_n	ρ_{τ_n}	$\frac{\widehat{\rho}^{*H}_{\tau_n}}{\rho_{\tau_n}}$	$\frac{\widehat{R}^{PL}_{\tau_n}}{\rho_{\tau_n}}$	$\frac{\widehat{R}^{Wins}_{\tau_n}}{\rho_{\tau_n}}$	MSE(H)	MSE(PL)	MSE(Wins)
(1000, 3, 30)	0.9700	32.7333	1.7438 (1.2840)	1.5512 (0.9589)	1.2495 (0.6447)	2.2003	1.2225	0.4775
(1000, 3, 40)	0.9600	32.7333	1.8345 (1.0530)	1.6454 (0.8650)	1.3740 (0.6160)	1.8040	1.1640	0.5190
(1000, 5, 30)	0.9700	6.4934	1.6003 (0.7143)	1.5128 (0.6155)	1.3441 (0.5052)	0.8701	0.6415	0.3733
(1000, 5, 40)	0.9600	6.4934	1.7371 (0.6666)	1.6367 (0.5828)	1.4855 (0.4981)	0.9872	0.7447	0.4835
(5000, 3, 150)	0.9700	32.7333	1.5032 (0.3798)	1.4148 (0.3479)	1.1596 (0.2410)	0.3974	0.2929	0.0835
(5000, 3, 200)	0.9600	32.7333	1.6273 (0.3875)	1.5175 (0.3431)	1.2808 (0.2543)	0.5435	0.3854	0.1435
(5000, 5, 150)	0.9700	6.4934	1.5021 (0.2644)	1.4459 (0.2405)	1.2913 (0.2014)	0.3220	0.2566	0.1253
(5000, 5, 200)	0.9600	6.4934	1.6372 (0.2763)	1.5659 (0.2473)	1.4251 (0.2147)	0.4822	0.3814	0.2268
(20000, 3, 600)	0.9700	32.7333	1.4411 (0.1749)	1.3641 (0.1601)	1.1260 (0.1142)	0.2251	0.1582	0.0289
(20000, 3, 800)	0.9600	32.7333	1.5772 (0.1833)	1.4764 (0.1623)	1.2531 (0.1228)	0.3667	0.2533	0.0791
(20000, 5, 600)	0.9700	6.4934	1.4663 (0.1266)	1.4152 (0.1162)	1.2669 (0.0973)	0.2335	0.1858	0.0807
(20000, 5, 800)	0.9600	6.4934	1.6094 (0.1346)	1.5419 (0.1214)	1.4060 (0.1053)	0.3895	0.3084	0.1759
(50000, 3, 1500)	0.9700	32.7333	1.4384 (0.1159)	1.3640 (0.1066)	1.1258 (0.0760)	0.2056	0.1439	0.0216
(50000, 3, 2000)	0.9600	32.7333	1.5792 (0.1178)	1.4798 (0.1057)	1.2558 (0.0793)	0.3493	0.2414	0.0717
(50000, 5, 1500)	0.9700	6.4934	1.4657 (0.0840)	1.4155 (0.0772)	1.2672 (0.0647)	0.2239	0.1786	0.0756
(50000, 5, 2000)	0.9600	6.4934	1.6124 (0.0866)	1.5452 (0.0784)	1.4090 (0.0679)	0.3825	0.3034	0.1719

Table 2.2: Based on the sample of size $n = 1000, 5000, 20000$ and 50000 from the Pareto$(1/\gamma)$ distribution with $\tau = 0.9995$, the true values of the extreme Wang distortion risk measure ρ_{τ_n} and the sample means, sample standard deviations of $\frac{\widehat{\rho}^{*H}_{\tau_n}}{\rho_{\tau_n}}, \frac{\widehat{R}^{PL}_{\tau_n}}{\rho_{\tau_n}}, \frac{\widehat{R}^{Wins}_{\tau_n}}{\rho_{\tau_n}}$ and MSEs are presented for various values of k.

2.4.3 Bias-corrected level

We simulate 1000 replications of random samples from the Pareto$(1/\gamma)$ distribution. All the experiments use sample sizes of $n = 1000, 5000, 20000$ and 50000 and the true index of the Pareto distribution $\gamma = 1/3, 1/5$. Moreover, we pick the $\alpha = 1.2$. The choices of γ and α satisfy $\alpha < 1/\gamma$. In Table 2.3, we report the mean, sample standard deviation and MSEs of the ratio of the estimator to its true value. In Figure 2.4, we plot the MSEs and the sample standard deviations against different q_n levels for samples of size 20000. Our estimator's accuracy is very close to the reduced bias version of the Wins-PL for small values of n, and it outperforms its results for large sample sizes.

2.5 TD as a Reinsurance Premium Principle

We define the following generalized tail distortion reinsurance premium principle

$$\rho_p^R(X) = \int_0^1 h\left(U\left(\frac{1}{q(1-p)}\right) - U\left(\frac{1}{(1-p)}\right)\right) dg(q).$$

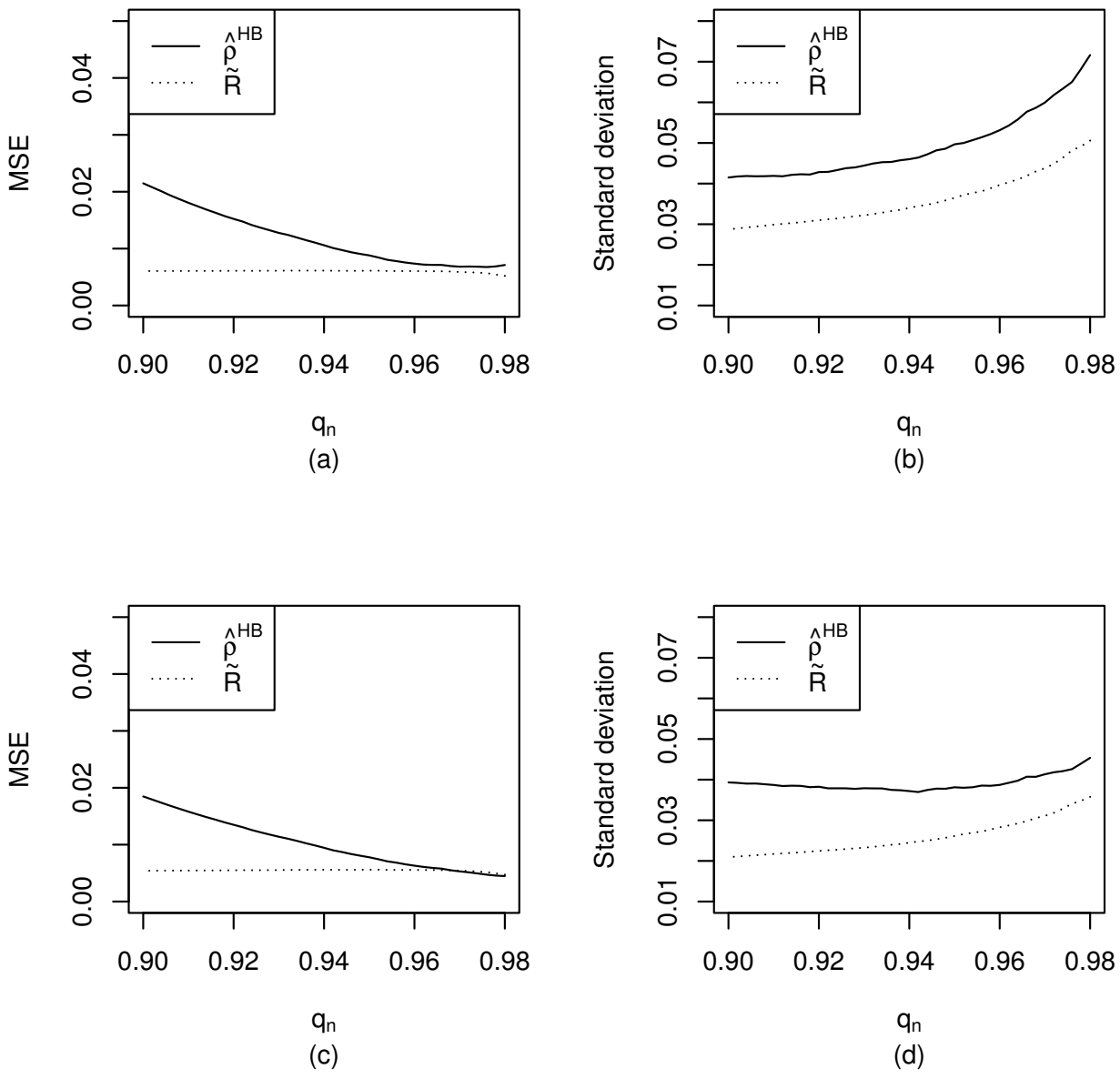


Figure 2.4: The MSEs (a,c) and sample standard deviations (b,d) of $\frac{\widehat{\rho}^{H}_{q_n}}{\rho_{q_n}}$ (solid line), $\frac{\widetilde{R}_{q_n}}{\rho_{q_n}}$ (dash-dotted line) are plotted against q_n based on 1000 repetitions sampling 20000 times from the Pareto(3) and Pareto(5). The (a) and (b) are MSE and sample standard deviations of Pareto (3) and the (c) and (d) are MSE and sample standard deviations of Pareto(5).

$(n, 1/\gamma, k)$	q_n	ρ_{q_n}	$\frac{\widehat{\rho}^{HB}_{q_n}}{\rho_{q_n}}$	$\frac{\widehat{\rho}^{H}_{q_n}}{\rho_{q_n}}$	MSE(HB)	MSE(H)
(1000, 3, 30)	0.9700	5.1921	0.3982 (14.4275)	1.0972 (0.2618)	208.3081	0.0779
(1000, 3, 40)	0.9600	4.4899	0.9735 (0.7792)	1.1006 (0.2184)	0.6072	0.0577
(1000, 5, 30)	0.9700	1.6819	0.9256 (0.6197)	1.0482 (0.1453)	0.3892	0.0234
(1000, 5, 40)	0.9600	1.4985	0.8940 (0.2061)	1.0547 (0.1301)	0.0537	0.0199
(5000, 3, 150)	0.9700	5.1921	0.9553 (0.1308)	1.0670 (0.1000)	0.0191	0.0145
(5000, 3, 200)	0.9600	4.4899	0.9446 (0.1150)	1.0760 (0.0915)	0.0163	0.0141
(5000, 5, 150)	0.9700	1.6819	0.9298 (0.0877)	1.0411 (0.0629)	0.0126	0.0056
(5000, 5, 200)	0.9600	1.4985	0.9200 (0.0822)	1.0475 (0.0579)	0.0132	0.0056
(20000, 3, 600)	0.9700	5.1921	0.9431 (0.0599)	1.0547 (0.0470)	0.0068	0.0052
(20000, 3, 800)	0.9600	4.4899	0.9326 (0.0531)	1.0665(0.0439)	0.0074	0.0064
(20000, 5, 600)	0.9700	1.6819	0.9400 (0.0413)	1.0351 (0.0300)	0.0053	0.0021
(20000, 5, 800)	0.9600	1.4985	0.9307 (0.0387)	1.0428 (0.0281)	0.0063	0.0026
(50000, 3, 1500)	0.9700	5.1921	0.9429 (0.0381)	1.0550 (0.0311)	0.0047	0.0040
(50000, 3, 2000)	0.9600	4.4899	0.9309 (0.0347)	1.0674 (0.0285)	0.0060	0.0054
(50000, 5, 1500)	0.9700	1.6819	0.9450 (0.0237)	1.0356 (0.0198)	0.0036	0.0017
(50000, 5, 2000)	0.9600	1.4985	0.9352 (0.0223)	1.0435 (0.0182)	0.0047	0.0022

Table 2.3: Based on the sample of size $n = 1000, 5000, 20000$ and 50000 from the Pareto$(1/\gamma)$ distribution, the true values of the extreme Wang distortion risk measure ρ_{q_n} and the sample means, sample standard deviations and MSEs of $\frac{\widehat{\rho}^{HB}_{q_n}}{\rho_{q_n}}, \frac{\widehat{\rho}^{H}_{q_n}}{\rho_{q_n}}$, are presented for various values of k.

When $g(x) = x$ and $h(x) = x^\alpha$, we have

$$\rho_p^R(X) = E[(X - \mathrm{VaR}_p(X))^\alpha \,|X > \mathrm{VaR}_p(X)].$$

Similar to Theorem 2.2.1, we can show the first-order expansion of $\rho_p^R(X)$ as follows.

Theorem 2.5.1. *Assume that $U \in \mathrm{RV}_\gamma$ with $\gamma > 0$ and $h \in \mathrm{RV}_\alpha$ with $\alpha > 0$. Further assume that $\int_1^\infty g\left(x^{-1/(\alpha\gamma+\delta)}\right) \mathrm{d}x < \infty$ for some $\delta > 0$. Then*

$$\rho_p^R[X] \sim \lambda_\gamma^R h\left(F^{\leftarrow}(p)\right), \tag{2.5.1}$$

where $\lambda_\gamma^R = \int_1^\infty g\left(\left(x^{1/\alpha}+1\right)^{-1/\gamma}\right) \mathrm{d}x$.

Proof. Note that

$$\lim_{p\to 1} \frac{h\left(U\left(\frac{1}{q(1-p)}\right) - U\left(\frac{1}{(1-p)}\right)\right)}{h\left(\gamma U\left(\frac{1}{(1-p)}\right)\right)} = \left(\frac{q^{-\gamma}-1}{\gamma}\right)^\alpha.$$

By following the similar arguments in Theorem 3.1 of Yang 2015 and by applying the

dominated convergence theorem, we can show that

$$\lim_{p\to 1}\int_0^1 \frac{h\left(U\left(\frac{1}{q(1-p)}\right)-U\left(\frac{1}{(1-p)}\right)\right)}{h\left(\gamma U\left(\frac{1}{(1-p)}\right)\right)}dg(q)=\int_0^1\left(\frac{q^{-\gamma}-1}{\gamma}\right)^{\alpha}dg(q).$$

Further noting that

$$h\left(\gamma U\left(\frac{1}{(1-p)}\right)\right)\sim\gamma^{\alpha}h\left(U\left(\frac{1}{(1-p)}\right)\right).$$

The desired result follows. □

The estimation of $\rho_p^R(X)$ at the intermediate level q_n is

$$\widehat{\rho}_{q_n}^R(X)=\widehat{\lambda}_{\gamma}^R h\left(X_{n-\lfloor n(1-q_n)\rfloor,n}\right)$$

where

$$\widehat{\lambda}_{\gamma}^R=\int_1^{\infty} g\left(\left(x^{1/\alpha}+1\right)^{-1/\widehat{\gamma}}\right)\mathrm{d}x.$$

The extreme level estimator at level τ_n is

$$\widehat{\rho}_{\tau_n}^{R*}(X)=\left(\frac{1-\tau_n}{1-q_n}\right)^{-\alpha\widehat{\gamma}}\widehat{\rho}_{q_n}^R(X).$$

The second-order expansion of $\rho_p^R(X)$ and the asymptotic normality of $\widehat{\rho}_{q_n}^R(X)$ and $\widehat{\rho}_{\tau_n}^{R*}(X)$ can be shown similarly as Theorems 2.2.2, 2.3.1, 2.3.2. For the sake of simplicity, we omit these results.

2.6 Application

In this section, we explore the use of the generalized tail distortion risk measure in practice. In order to compare with results from other research studies and other risk measures, we

choose $g(x) = x$ and $h(t) = t^\alpha$ for some $\alpha > 0$. We also compare the estimates of the risk measure with the VaR and ES for single risks, as such:

- compare single risk with VaR
$$\frac{\rho_p(X)}{\mathrm{VaR}_p(X)}$$
- compare single risk with ES
$$\frac{\rho_p(X)}{\mathrm{ES}_p(X)}$$

2.6.1 Analysis of automobile claims, the Secura Belgian Re actuarial data set

The Secura Belgian Re data set consists of 371 automobile claims from 1988 until 2011, collected from several European insurance companies. It was first introduced in Beirlant et al. (2004) and further studied in Vandewalle and Beirlant (2006) and El Methni and Stupfler (2017) for reinsurance premium pricing. All claim amounts were corrected for inflation and the smallest amount is 1.2 million euros. The data set was originally analyzed to provide the participating reinsurance companies with statistical tools to assist in pricing the unlimited excess-loss layer above a high retention level R. We aim to re-examine the data set and show that our estimator can provide results that are comparable to those explored in the literature.

We choose k from the first stable region of the Hill plot and accordingly estimate the tail index γ by the Hill estimator, shown in Figure 2.5. Thus, we have that $q_n = 0.830$ and $\widehat{\gamma} = 0.280$. Estimates of the tail distortion risk measure for the intermediate level $\widehat{\rho}^H_{q_n}(X)$ are shown in Table 2.4, calculated for several values of α. We also estimate the risk measure $\widehat{\rho}^{*H}_{\tau_n}(X)$ for several extreme levels $\tau_n = 0.98, 0.99, 0.995, 0.999$. Finally, we compare the estimates of $\widehat{\rho}^H_{q_n}(X)$ and $\widehat{\rho}^{*H}_{\tau_n}(X)$ to two of the most well-known risk measures, the VaR and the ES. From the results, we can confirm numerically that when $g(x) = x$

and $h(x) = x^{\alpha}$, for $\alpha = 1$, the tail distortion risk measure becomes equivalent to the ES. However, as the value of τ_n increases, $\widehat{\rho}^{*H}_{\tau_n}(X)$ increases more than $\mathrm{ES}_{\tau_n}(X)$, which is explained by the fact that the empirical estimators of the traditional risk measures are not consistent at extreme levels when the chosen level exceeds the range covered by the available data.

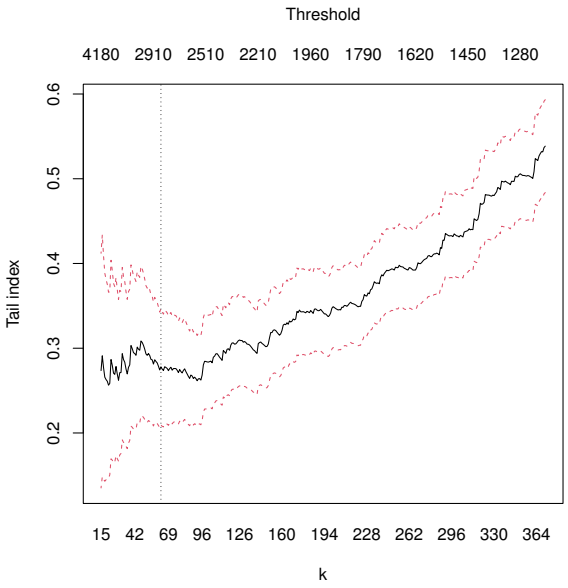


Figure 2.5: The Hill plot for the Secura Belgian Re data set. The red lines are the 95% confidence bands, the vertical dashed line is our chosen k and the vertical dotted line corresponds to the k used in El Methni and Stupfler (2017). Threshold values are in thousands of Euros.

The purpose of the analysis in Vandewalle and Beirlant (2006) and El Methni and Stupfler (2017) was to calculate the reinsurance premium for an excess-of-loss reinsurance policy by considering the net premium principle (NP). For $g(x) = x$ and $h(t) = t$, one has that

$$NP\left(\mathrm{VaR}_p(X)\right) = (1-p)\left(\rho_p(X) - \mathrm{VaR}_p(X)\right).$$

El Methni and Stupfler (2017) estimated the tail index by the median of the Hill estimator and some of its bias-reduced versions, yielding $\widehat{\gamma} = 0.261$ and $q_n = 0.792$. Estimates of $\widehat{\rho}^{*H}_{\tau_n}(X)$ at this risk level are provided in Table 2.5, indicating slightly higher estimates for

α	τ_n	$\widehat{\rho}^{H}_{q_n}(X)$	$\frac{\widehat{\rho}^{H}_{q_n}(X)}{\text{VaR}_{q_n}(X)}$	$\frac{\widehat{\rho}^{H}_{q_n}(X)}{\text{ES}_{q_n}(X)}$	$\widehat{\rho}^{*H}_{\tau_n}(X)$	$\frac{\widehat{\rho}^{*H}_{\tau_n}(X)}{\text{VaR}_{\tau_n}(X)}$	$\frac{\widehat{\rho}^{*H}_{\tau_n}(X)}{\text{ES}_{\tau_n}(X)}$
1	0.98	3,797	1.32	0.98	6907	1.31	1.04
	0.99				8,384	1.24	1.12
	0.995				10,179	1.37	1.32
	0.999				15,967	2.06	2.02
1.2	0.98	19,958	6.97	5.07	40,921	7.78	6.18
	0.99				51,641	7.64	6.95
	0.995				65,169	8.80	8.47
	0.999				111,857	14.43	14.16
1.4	0.98	105,425	36.82	26.77	243,629	46.34	36.84
	0.99				319,608	47.29	43.04
	0.995				419,281	56.62	54.50
	0.999				787,467	101.65	99.69

Table 2.4: Secura Belgian Re data set: Tail distortion risk measure estimated at $q_n = 0.83$ for the intermediate level and at $\tau_n = 0.98, 0.99, 0.995, 0.999$ for the extreme level, by using the estimators in Eqs. (2.3.2) and (2.3.5), respectively. The estimates are calculated for $\alpha = 1, 1.2, 1.4$ and $q_n = 0.830$, and are compared against the VaR and ES for single risks. Estimates are in thousands of Euros.

$\widehat{\rho}^{*H}_{\tau_n}(X)$ compared to $\widehat{R}^{PL}_{\tau_n}$, especially for larger values of τ_n. The estimates are compared against the ones calculated by El Methni and Stupfler (2017). The values are smaller than those provided in Table 2.4, which is explained by the smaller choices of q_n and $\widehat{\gamma}$.

Set the retention level $R = 5$ million Euros, which corresponds to $\text{VaR}_{0.98}(X)$ (i.e. $\tau_n = p = 0.98$). The reinsurance premium for the excess-of-loss reinsurance policy is 38,142 Euros (if γ is estimated by the Hill estimator) or 36,764 (if γ is estimated by a bias-reduced version of the Hill estimator). Those results are in line with the 36,000 Euros and 41,798 that El Methni and Stupfler (2017) and Vandewalle and Beirlant (2006) estimated, respectively. The smaller net premium values are due to the using the bias-reduced estimate of the tail index γ compared to the Hill estimator.

τ_n	$\widehat{\rho}^{*H}_{\tau_n}(X)$	$\widehat{R}^{PL}_{\tau_n}$ (El Methni and Stupfler 2017)
0.98	6,838	6,864
0.99	8,293	8,224
0.995	10,059	9,854
0.999	15,745	14,993

Table 2.5: Secura Belgian Re data set: Comparison between the Tail distortion risk measure $\widehat{\rho}^{*H}_{\tau_n}(X)$ and El Methni and Stupfler (2017)'s $\widehat{R}^{PL}_{\tau_n}$ estimated for $\alpha = 1$ at $q_n = 0.792$ for the extreme levels $\tau_n = 0.98, 0.99, 0.995, 0.999$.

When $g(x) = x$ and $h(x) = x^{\alpha}$, we have following reinsurance premium principles

- $\Pi_1 = (1-p)(\rho_{p,1}(X) - VaR_p(X)) = (1-p)\rho^R_{p,1}(X)$.
- $\Pi_2 = (1-p)\left(\rho_{p,\alpha}(X) - VaR_p(X)\right) = (1-p)\left(E\left[X^\alpha | X > VaR_p(X)\right] - VaR_p(X)\right)$
- $\Pi_3 = (1-p)\rho^R_{p,\alpha}(X) = (1-p)E\left[\left(X - VaR_p(X)\right)^\alpha | X > VaR_p(X)\right]$

For the above three premium principles, we divide the data set into a training and testing data sets, composed of 300 and 71 observations, respectively. The training data set is used to estimate γ and calculate the reinsurance premium, based on multiple retention levels. Table 2.6 summarizes the average of the reinsurance premium values over 100 cross-validation repetitions of the analysis. For $\alpha = 1$, Π_1 is equivalent to Π_3. Given the reinsurance arrangement, the average premium amount decreases as the retention level increases.

α	p	Π_1	Π_2	Π_3
1	0.950	59,965	61,994	59,965
	0.990	17,614	22,253	17,614
	0.995	10,392	6,321	10,392
1.2	0.950		1,205,431	284,056
	0.990		390,861	90,107
	0.995		232,593	54,954

Table 2.6: Secura Belgian Re data set: Reinsurance premium based on 100 cross-validation repetitions of sampling of 300 observations.

With a reinsurance arrangement, the insured loss under a premium principle Π_i for the testing data sets is $\sum_j (\Pi_i + \min\{X_j, R\})$, for $i = 1, \ldots, 3$, and $j = 1, \ldots, 71$. While without reinsurance, the uninsured loss is total loss amounts $\sum_j X$, for $j = 1, \ldots, 71$. Table 2.7 provides the summary statistics of the insured and unisured losses, based on the testing samples of 71 observations. The insured losses have lower variability, and hence offer better management of the tail risk.

α	(Un)Insured	p	Mean	Standard deviation	$\text{VaR}_{0.9}$	$\text{VaR}_{0.95}$	$\text{VaR}_{0.99}$
	Uninsured		2,391	1,248	4,017	5,446	6,757
1	Π_1	0.950	2,266	765	3,841	3,841	3,841
		0.990	2,347	1,060	4,035	5,167	5,167
		0.995	2,401	1,248	4,028	5,456	6,767
	Π_2	0.950	2,268	765	3,843	3,843	3,843
		0.990	2,351	1,060	4,040	5,172	5,172
		0.995	2,397	1,248	4,024	5,452	6,763
	Π_3	0.950	2,266	765	3,841	3,841	3,841
		0.990	2,347	1,060	4,035	5,167	5,167
		0.995	2,401	1,248	4,028	5,456	6,767
1.2	Π_2	0.950	3,411	765	4,987	4,987	4,987
		0.990	2,720	1,060	4,408	5,540	5,540
		0.995	2,623	1,248	4,250	5,678	6,989
	Π_3	0.950	2,490	765	4,065	4,065	4,065
		0.990	2,419	1,060	4,108	5,240	5,240
		0.995	2,446	1,248	4,072	5,500	6,812

Table 2.7: Secura Belgian Re data set: Summary statistics of the insured and uninsured losses, based on 100 cross-validation testing samples of 71 observations. Estimates are in thousands of Euros.

2.7 Conclusion

In this paper, we have focused on the generalized tail distortion risk measure for extreme risks, which is used to assess tail risks of excess losses modeled by the right tails of loss distributions. We derive an estimator for the risk measure by using its asymptotic expansions. This estimator is asymptotically normal at the intermediate and extreme confidence levels. Additionally, by using the Hill estimator, we provide bias-corrected estimators. We perform a simulation study to assess the accuracy of the estimator over a range of confidence levels and distortion factor α, in which we confirm that our estimator provides good results, especially for larger sample sizes. Finally, we conclude this paper by illustrating an application on a real-life actuarial data set and we compare our estimators to the traditional risk measures, VaR and ES, which, unlike our estimator, their empirical estimators are not consistent at extreme levels when the chosen confidence level exceeds the range covered by the observed data. We also propose a reinsurance premium principle based on the generalized tail distortion risk, which provides better management of the tail risk for an insurance company.

Chapter 3

Goodness-of-fit tests for Space-time Point Process Models: An Application to Earthquake Insurance Risk

3.1 Introduction

Catastrophic losses from earthquakes occurring in densely populated areas may cause a serious threat to the financial and economic stability of Property and Casualty (P&C) insurance and reinsurance companies. Earthquakes happen randomly and without any reliable means of predicting their exact location or time of occurrence (U.S. Geological Survey 2022a). Unlike most meteorological disasters, earthquakes have very long return periods and accordingly small losses in recent years may not be indicative of future losses. Around half a million earthquakes are detected annually around the globe, with only 100 of them being strong enough to cause damage (U.S. Geological Survey 2021). The infrastructure of buildings, their design and material are major factors that explain the extent of damage caused by an earthquake and its subsequent events.

In Canada, earthquakes occur every year at a very high frequency, but on average, only

one earthquake per week is large enough to be felt by residents (NRC 2019). Earthquakes that are strong enough to cause material damage occur decades apart and they usually strike offshore or in locations that are not populated. Canada's west coast falls in the Circum-Pacific seismic belt, also known as the "the Ring of Fire", where 81% of the world's largest earthquakes occur (U.S. Geological Survey 2022b). On the other hand, seismic activity in Eastern Canada is attributed to regional stress fields and strong earthquakes occur at a relatively lower rate.

According to the United States Geological Survey, earthquakes in Eastern North America can be felt at much further distances than earthquakes in Western North America of comparable size. Geologists attribute such differences to the nature of the underlying tectonic plates in the regions, and the size and age of buildings. Eastern North America has rocks that have been formed billions of years before the rocks in Western North America. These old rock formations are currently harder and denser due to their exposure to extreme pressures and temperatures and the faults have had more time to heal. Hence, seismic waves travel longer, compared to the younger faults in the west which absorb a lot of the seismic wave energy and minimize its spread (U.S. Geological Survey 2018).

Figure 3.1 provides a spatial representation of the historical seismicity of significant earthquakes in Canada, for a total of 172 events for the period $1600 - 2017$ (Lamontagne et al. 2018). An earthquake is considered significant if its moment magnitude exceeded 6, and/or it had been reported as felt by residents at Modified Mercalli intensity (MMI) of V or higher. The definitions of the MMI levels are available in Appendix I.1. Given the actuarial nature of our study, we are interested only in significant earthquakes because claims are more likely to occur due to them. The magnitude of an earthquake is a value that describes its size, whereas the intensity is a measure of shaking, which varies across locations, depending mostly on nearness to the epicenter. The MMI level assigned to a location following an earthquake provides a more interpretable measure to

non-seismologists than the magnitude because it quantifies the felt shaking. Initially, the MMI was a subjective measure approximated from humans' reports of felt ground shaking, but now, instrumental intensity measures are used to estimate the MMI, such as the ground motion-intensity relationship derived by Wald et al. (1999).

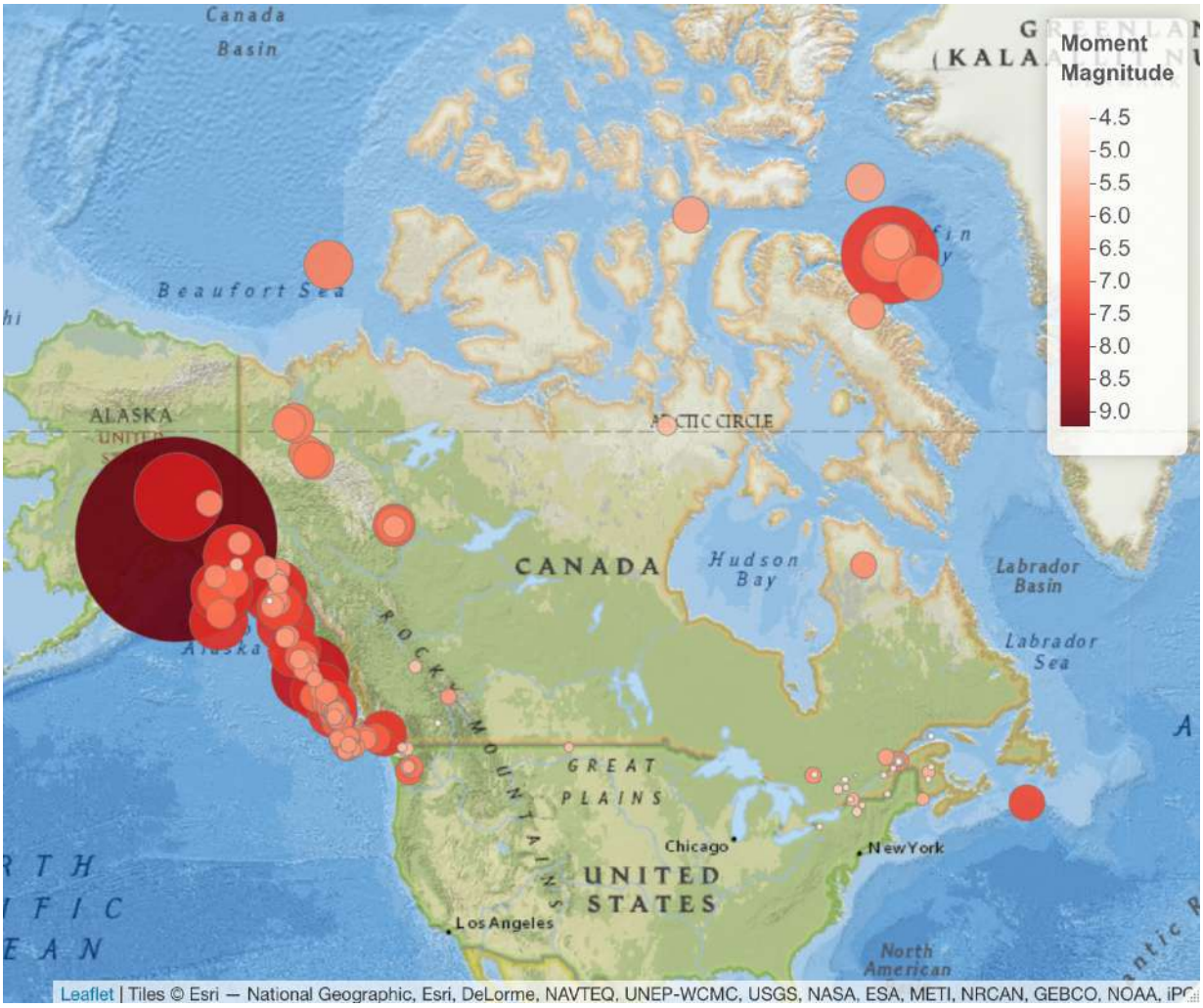


Figure 3.1: Significant Canadian earthquakes for the period 1600 – 2017. The size and color of the circles are proportionate to the moment magnitude.

Earthquake losses are not covered by a standard home insurance policy. Earthquake insurance usually applies a large deductible, and accordingly, the loss amount has to be substantially high before the insurance payments can be made, which may discourage homeowners from purchasing earthquake insurance coverage. British Columbia (BC) and

the Ontario (ON) and Québec (QC) region are at risk due to their large population density and elevated level of seismic activity. More than 70% of the population in QC resides in seismic regions, yet only 3.4% of homeowners in the cities of Montréal and Québec City hold earthquake insurance, compared to 65% market penetration in the cities of Vancouver and Victoria (Swiss Re 2017).

The Office of the Superintendent of Financial Institutions (OSFI) is the Canadian organisation which sets guidelines for the recommended assessment of insurers' catastrophe risk management and calculation of probable maximum loss (PML), the threshold dollar value of earthquake losses beyond which losses are unlikely. The insurance risk arising from earthquakes is one of the components of the Minimum Capital Test (MCT) for federally regulated P&C insurance companies. The formula for the calculation of the earthquake reserve as set by OSFI (2019) is

$$\text{Country-wide PML}_{1/500} = \left(\text{East Canada PML}_{1/500}^{1.5} + \text{West Canada PML}_{1/500}^{1.5}\right)^{\frac{1}{1.5}}, \quad (3.1.1)$$

where the earthquake PML is the gross PML, which is the PML amount after deductibles but before any reinsurance deductions. $\text{PML}_{1/500}$ is a 1-in-500 year event, representing the 99.8^{th} percentile of the distribution of the annual maxima.

Due to the complexity of modeling seismic activity, Canadian insurers are expected to use earthquake models to calculate their earthquake reserve. AIR prepared an earthquake model for the insurance industry that provides estimates of the damages and economic losses due to all earthquake-related events for Canada. Another Canadian earthquake model is HazCan (Ulmi et al. 2014), the Canadian version of HazUS (FEMA 2013). These models provide a comprehensive study of probabilistic loss scenarios by relying on an inventory of population and building information as well as ground motion equations, but they are usually complex, sold at a high price, not available for public use and difficult to replicate.

Spatio-temporal modeling of earthquakes has witnessed considerable progress, especially in seismically active regions in North America. Such models include the Regional Earthquake Likelihood Models (RELM) project (Field 2007) and the Collaboratory for the Study of Earthquake Predictability (CSEP) (Jordan 2006), which focus on the known faults in California. To assess the fitted models in the RELM project, Schorlemmer et al. (2007) develop several numerical tests, such as the Number-test (or N-test), the Likelihood-test (or L-test) and the Likelihood-ratio-test (R-test). The N-test compares the total estimated and observed count of events, the L-test assesses the overall likelihood of the fitted model, and the R-test compares the relative performance of two fitted models. Those methods are successful in assessing the overall fit of the model, however, they are incapable of identifying locations or time periods where the fit is poor. Recent progress in measuring the goodness-of-fit and comparing spatio-temporal models include pixel-based residual methods, in which the spatial region is divided into a predetermined regular grid and a residual is computed for each pixel, see Baddeley et al. (2005) and Zhuang (2006). Pixel-based residual analysis, such as raw residuals and Pearson residuals, show directly where the fitted models may be improved, while deviance residuals rank the performance of two fitted models. For models where the expected number of events at a pixel is close to zero, some problems may arise because of the heavy skewness of the distribution of the residual for the given pixel. To overcome these deficiencies, Bray et al. (2014) use Voronoi polygons, generated by the observed spatio-temporal point pattern, which are adaptive to the inhomogeneity of the process, to compare the expected and observed count of events.

Bray et al. (2014) noted that the standardization techniques and other residual methods studied by Baddeley et al. (2005) and Clements, Schoenberg, and Veen (2012) may be applied to Voronoi polygons. In this research project, we implement this note by extending the Voronoi residual methods, i.e. we calculate the Pearson Voronoi residuals, in analogy with Pearson residuals for generalized linear models. To compare competing models, we extend the definition of pixel-based deviance residuals to Voronoi polygons.

In analogy with linear models, the resulting residuals may be called deviance Voronoi residuals. We use them to assess the goodness-of-fit of fitted spatio-temporal models for the significant Canadian earthquakes point pattern. Having established a well-fitting model, we analyze earthquake insurance risk by creating an open-source and reproducible simulation-based approach. An interactive web application allowing the user to simulate a significant earthquake based on any chosen location is also provided. Additionally, we review OSFI's MCT formula in eq. (3.1.1) and provide a possible alternative.

The remainder of the paper is organized as follows. Section 3.2 provides a brief summary of spatio-temporal point processes, some of their properties and goodness-of-fit tests to assess the adequacy of fitted spatio-temporal models. Section 3.3 explains the methodology used to calculate the building and content exposure, simulate a location and year of occurrence of an earthquake and estimate its financial impact. It also proposes an alternative approach to calculate the MCT for earthquake risk in Canada. Section 3.4 compares our results with other results from simulated earthquakes in the literature, summarizes the outcome of our methodology and compares OSFI's MCT approach for earthquake risk to our alternative proposal. Section 3.5 concludes the article. The unit of currency throughout is the Canadian dollar, unless otherwise stated.

3.2 STPP and their Residuals

A Spatio-Temporal Point Process (STPP) is a stochastic process that models data of the location and time of occurrence of events. Examples of processes that can be modeled include spread of diseases and pandemics, natural disasters such as earthquakes, tsunamis and volcanic eruptions. Point processes are studied thoroughly in time; see Cox and Isham (1980) and Daley and Vere-Jones (2003) and in space; see Cressie (2015), Moller and Waagepetersen (2003), and Diggle (2013). STPP are extensively used in seismology, but

mainly to study earthquake aftershocks; see Ogata (1998) and Zhuang, Ogata, and Vere-Jones (2002).

A STPP is defined as a random measure on a region $S \subseteq \mathbb{R}^2 \times \mathbb{R}^+$ of space-time, where the spatial component is located in two spatial coordinates: longitude and latitude. A realization of a STPP is a spatio-temporal point pattern (stpp) consisting of location $\boldsymbol{x}_i$ (longitude and latitude) and a corresponding time of occurrence t_i for an event, such that $\{(\boldsymbol{x}_i, t_i) : i = 1, \ldots, n\}$, where $(\boldsymbol{x}_i, t_i) \in A \times T$ for some known spatial region A and temporal period T.

The first-order properties of a STPP are defined by its spatio-temporal intensity function given by

$$\lambda(\boldsymbol{x}, t) = \lim_{|d\boldsymbol{x}|,|dt|\to 0} \frac{\mathbb{E}\left[N(d\boldsymbol{x}, dt)\right]}{|d\boldsymbol{x}||dt|},$$

where $d\boldsymbol{x}$ represents a small spatial region around the location x such that $|d\boldsymbol{x}|$ is its area, dt represents a small time interval containing the time point t such that $|dt|$ is its length and $N(d\boldsymbol{x}, dt)$ represents the number of events in $d\boldsymbol{x} \times dt$. Thus, $\lambda(\boldsymbol{x}, t)$ represents the mean number of events per unit area per unit of time. For a homogeneous STPP, $\lambda(\boldsymbol{x}, t) = \lambda$ for all $(\boldsymbol{x}, t) \in A \times T$. In practice, the STPP is observed on a finite space-time region $A \times T$, such that the marginal spatial and temporal intensities can be defined as

$$\lambda_A(\boldsymbol{x}) = \int_T \lambda(\boldsymbol{x}, t)dt, \quad \text{and} \quad \lambda_T(t) = \int_A \lambda(\boldsymbol{x}, t)d\boldsymbol{x},$$

respectively.

The N-test and the L-test are used to assess the adequacy of the fitted models (Schorlemmer et al. 2007). For a fixed number s of simulated realizations from the fitted model, the N-test requires computation of the fraction of simulations where the number of simulated events is less than the number of observed events. If the fraction is close to 0 or 1, the model is rejected. Similarly, the L-test requires computation of the fraction of simulated

log-likelihoods that are less than the log-likelihood of the observed stpp. If the fraction is close to 0, the model is rejected. The two methods only consider the overall fit of the model and do not identify locations or time periods where the fit is poor.

To overcome these deficiencies, Baddeley et al. (2005) introduce pixel-based residual analysis methods for spatial point processes and Zhuang (2006) extends to spatio-temporal point processes. Consider a STPP with conditional intensity $\hat{\lambda}(\boldsymbol{x}, t)$ at any location $\boldsymbol{x}$ and time of occurrence t. The simplest residual form, namely the raw residual, compares the total number of observed and estimated points within evenly spaced pixels over a regular grid, i.e.

$$r^R(B_i) = N(B_i) - \int_{B_i} \hat{\lambda}(\boldsymbol{x}, t) dt d\boldsymbol{x},$$

where $N(B_i)$ is the number of points in bin B_i, which is a measurable set such that $B_i \subset A \times T$, for $i = 1, \ldots, m$, where m is the total number of pre-determined bins. Thus, we can evaluate locations and time periods where the fitted intensity function fits poorly compared to the observed stpp. The residuals within each pixel can be standardized in numerous ways, which is important because otherwise residual plots can exaggerate the deviation in the pixels where the raw residual value is large. For example, the Pearson residuals are re-scaled raw residuals that have mean 0 and variance approximately equal to 1, in analogy with Pearson residuals in linear models, i.e.

$$r^P(B_i) = \sum_{(t_j, \boldsymbol{x}_j) \in B_i} \frac{1}{\sqrt{\hat{\lambda}(\boldsymbol{x}_j, t_j)}} - \int_{B_i} \sqrt{\hat{\lambda}(\boldsymbol{x}, t)} dt d\boldsymbol{x},$$

for all $\hat{\lambda}(\boldsymbol{x}_i, t_i) > 0$. See Baddeley et al. (2005) for other methods of scaling the gridded residuals and Baddeley, Møller, and Pakes (2008) for their properties. Analogous to deviances for regression models, Wong and Schoenberg (2009) propose using deviance residuals to differentiate model fits. Given two fitted intensity functions, $\hat{\lambda}_1$ and $\hat{\lambda}_2$, the

deviance residual for pixel B_i is the difference between their log-likelihoods, i.e.

$$r^D(B_i) = \left(\sum_{(t_j, \boldsymbol{x}_j) \in B_i} \log(\hat{\lambda}_1(\boldsymbol{x}_j, t_j)) - \int_{B_i} \hat{\lambda}_1(\boldsymbol{x}, t) dt d\boldsymbol{x} \right) - \left(\sum_{(t_j, \boldsymbol{x}_j) \in B_i} \log(\hat{\lambda}_2(\boldsymbol{x}_j, t_j)) - \int_{B_i} \hat{\lambda}_2(\boldsymbol{x}, t) dt d\boldsymbol{x} \right).$$

Despite the mentioned advantages of pixel-based residuals, they can be problematic when the expected number of points at a pixel, based on a fitted model, is very small, causing heavy skewness of the distribution of its residual. This is problematic for earthquake occurrences modeling because the expected counts of events that are far from previous seismicity are usually zero. A possible solution could be to increase the size of the pixels such that the expected number of points in each pixel is large enough. However, this can be problematic for inhomogeneous processes because it is difficult, or almost impossible, to identify a one-size-fits-all pixel size. Accordingly, Bray et al. (2014) propose Voronoi tessellations, which are adaptive to the inhomogeneity of the process. Consider a stpp consisting of location $\boldsymbol{x}_i$ and time t_i for an event, such that $\{(\boldsymbol{x}_i, t_i) : i = 1, \ldots, n\}$. A Voronoi tessellation is a partitioning of the spatial plane of interest into n convex polygons, known as Voronoi polygons. Each polygon C_i consists of the region of all locations on the spatial grid that are closer to $\boldsymbol{x}_i$ than to any other point in the process, for all $i = 1, 2, \ldots, n$. Figure 3.2 represents the Voronoi tessellations from Figure 3.1. For more details on Voronoi tessellations and their properties, see Boots, Okabe, and Sugihara (1999). Given a fitted intensity model, $\widehat{\lambda}(\boldsymbol{x}, t)$, raw Voronoi residuals are calculated by evaluating the raw residuals over the Voronoi polygons instead of the pixels. By definition, each Voronoi polygon contains exactly one observed point, i.e. $N(C_i) = 1, \forall i = 1, 2, \ldots, n$. Accordingly, the raw Voronoi residuals are defined by

$$r_V^R(C_i) = N(C_i) - \int_{C_i} \hat{\lambda}(\boldsymbol{x}, t) dt d\boldsymbol{x},$$

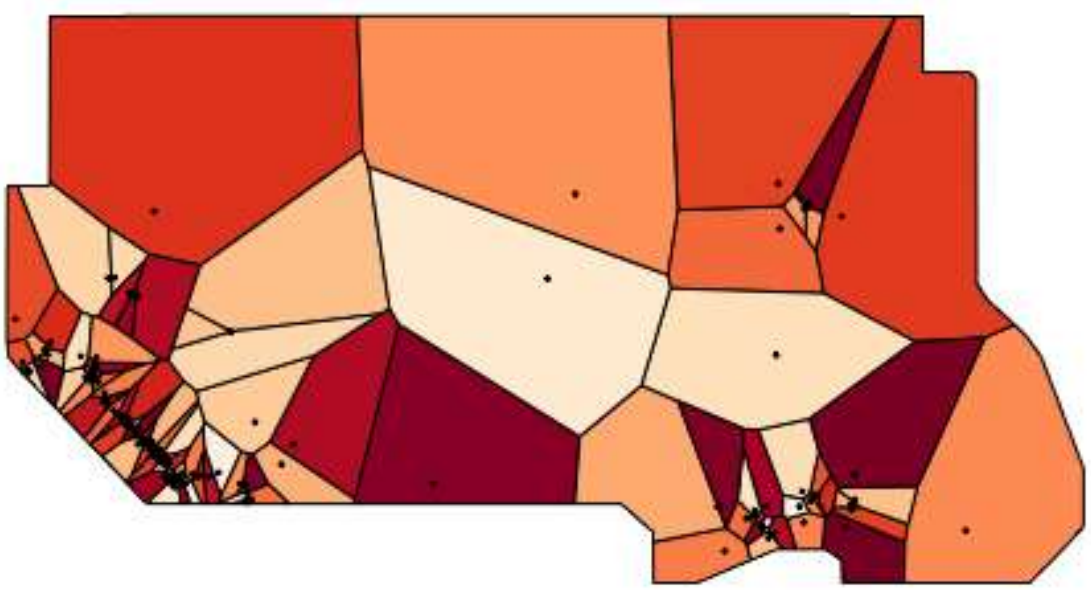


Figure 3.2: Voronoi tessellation of the significant Canadian earthquakes, shown in Figure 3.1.

$$= 1 - \int_{C_i} \hat{\lambda}(\boldsymbol{x}, t) dt d\boldsymbol{x}.$$

Voronoi residual plots are constructed by shading each polygon according to the residual value under the distribution function of the modified gamma distribution, which is the approximate distribution of the raw residuals (Bray et al. 2014). Gordon et al. (2015) use the Voronoi raw residuals of a homogeneous Poisson process model as a benchmark or scale to assess the Voronoi raw residuals plots of alternative models. It was noted by Bray et al. (2014) that the standardization techniques may be applied to Voronoi polygons.

In this research project, we implement this approach of extending the Voronoi residual methods. Accordingly, we scale the raw Voronoi residuals to obtain the Pearson Voronoi residuals, defined by

$$r_V^P(C_i) = \frac{1}{\sqrt{\hat{\lambda}(\boldsymbol{x}_i, t_i)}} - \int_{C_i} \sqrt{\hat{\lambda}(\boldsymbol{x}, t)} dt d\boldsymbol{x}.$$

The fitted intensity must satisfy $\hat{\lambda}(\boldsymbol{x}_i, t_i) > 0$ for all $\boldsymbol{x}_i$ and t_i in order that the Pearson Voronoi residuals be well-defined. Accordingly, this will yield residuals that have mean 0 and variance approximately equal to 1. Additionally, to compare competing models, we extend the definition of pixel-based deviance residuals to Voronoi Polygons, i.e. to calculate, in each Voronoi polygon, the difference between the log-likelihood for the two models. In analogy with linear models, the resulting residuals may be called deviance Voronoi residuals, defined by

$$r_V^D(C_i) = \left(\log(\hat{\lambda}_1(\boldsymbol{x}_i, t_i)) - \int_{C_i} \hat{\lambda}_1(\boldsymbol{x}, t) dt d\boldsymbol{x}\right) - \left(\log(\hat{\lambda}_2(\boldsymbol{x}_i, t_i)) - \int_{C_i} \hat{\lambda}_2(\boldsymbol{x}, t) dt d\boldsymbol{x}\right) \tag{3.2.1}$$

Accordingly, we can see, for each polygon, where one model outperforms the other. Polygons with a positive (negative) deviance Voronoi residual indicate that $\hat{\lambda}_1$ ($\hat{\lambda}_2$) provides a better fit. Summing the deviance residuals, i.e. $\sum_i r_V^D(C_i)$ yields a log-likelihood ratio score. This score provides a general indication of the improvement in fit provided by the best fitting model. Should one have more than two models, they can be compared two at a time to rule out the worst models. In Section 3.3, we apply the methods above to examine and compare several fitted spatio-temporal models to the significant Canadian earthquakes stpp.

3.3 Earthquake Simulation Methodology

This section explains the methodology to calculate the building exposure and to simulate earthquake financial losses. We also propose an alternative formula to eq. (3.1.1) for the MCT for earthquake risk in P&C insurance companies.

3.3.1 Calculating exposure and simulating earthquake damage

To estimate the seismic risk for a study area, we proceed with the following steps:

(i) Collecting building inventory and calculating exposure,

(ii) Simulating earthquakes in space and time,

(iii) Estimating ground shaking intensity,

(iv) Calculating damage rates, and

(v) Estimating seismic loss and insurance claim payments.

Parts (ii) - (v) of the methodology are summarized in Algorithms 1 and 2. To improve precision, the resolution of the analysis is at the Census Subdivisions (CSD) level, the municipalities used for statistical reporting purposes at Statistics Canada (Statistics Canada, Government of Canada 2017). There are 5,162 CSD in Canada.

(i) Collecting building inventory and calculating exposure

For residential buildings, we calculate the total square footage by using the number of buildings in each residential building classification from the 2016 Canadian Census (Statistics Canada, Government of Canada 2016) and the average square footage for each residential building from the Canadian Housing Statistics Program (Statistics Canada, Government of Canada 2020a). For each building type, the building exposure is calculated by multiplying the total square footage by the building replacement cost, in dollar units. The assumptions for replacement costs are taken from HazCan (Ulmi et al. 2014). The building construction price index (BCPI) provided by Statistics Canada, Government of Canada (2020b) is used to inflate the construction costs to June 2021. To account for regional and embedded arbitrariness between construction

companies, a random noise of ±10% is factored in the replacement cost. The building content replacement value for residential dwellings is assumed to be at 50% of the building replacement cost (Ulmi et al. 2014; FEMA 2013). The building exposure based on the building types is then distributed based on the building construction material (wood, concrete, steel, masonry, etc.) by using HAZUS' general building scheme mapping information provided in Table 5.1 in FEMA (2013).

Calculating the exposure of non-residential dwellings is not as straightforward because Statistics Canada currently does not have a comprehensive dataset for non-residential buildings. Accordingly, we rely on building permits data from Statistics Canada, Government of Canada (2020c) and use the annual ratios of institutional and governmental, commercial and industrial building permits to residential building permits. Hence, the non-residential buildings exposure is calculated as a percentage of the residential buildings exposure. Table 6.2 in Appendix I.2 provides the suggested conversion of non-residential building types to HAZUS occupancy codes. As with residential buildings, the exposure is classified based on the construction material (wood, concrete, steel, masonry, etc.) by using HAZUS' general building scheme mapping information, provided in Table 5.1 in FEMA (2013). The building content replacement value for non-residential dwellings is assumed to be a percentage of the building replacement cost as suggested in FEMA (2013).

Figure 3.3 summarizes the computed residential and non-residential exposure values per province. The values on the y-axis are omitted to maintain the confidentiality of CatIQ's (Canada's Loss and Exposure Indices Provider) exposure data, see Section 3.4.1 for details. Western Canada includes British Columbia (BC), Alberta (AB), Saskatchewan (SK), Manitoba (MB), Northwest Territories (NT) and Yukon (YT), while Eastern Canada includes Newfoundland and Labrador (NL), Nova Scotia (NS), Prince Edward Island (PE), New Brunswick (NB), Québec (QC), Ontario (ON) and Nunavut (NU).

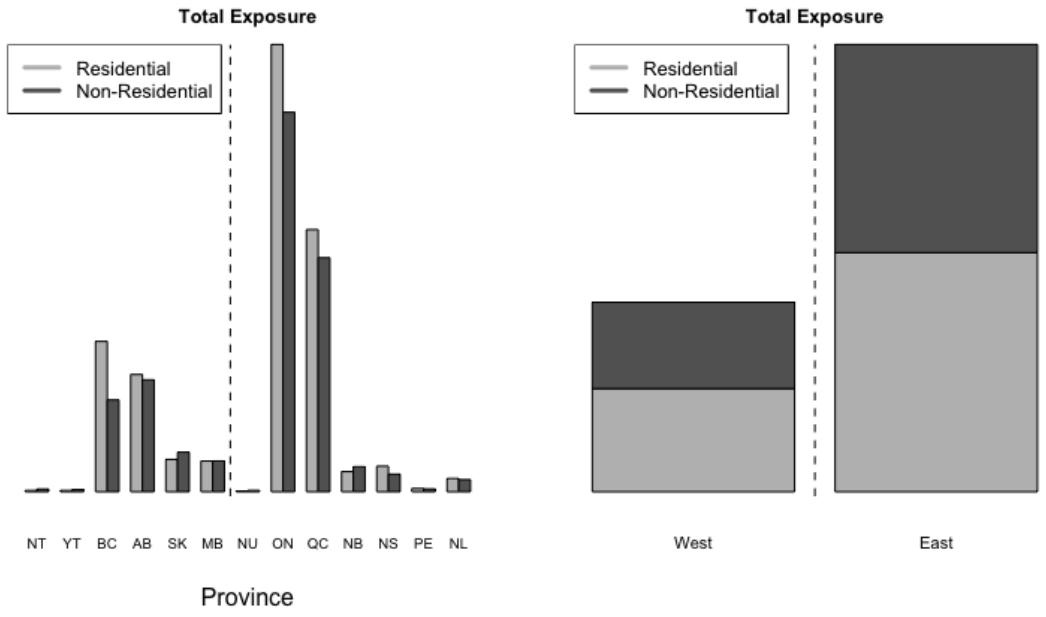


Figure 3.3: Total exposure per province (left) and for Eastern and Western Canada (right). The vertical dashed line splits Eastern and Western provinces.

Figure 3.4 provides a spatial representation of our calculated total exposure, including residential and non-residential building exposure and building contents exposure, per km^2 for each CSD.

(ii) Simulating earthquakes in space and time

By using the geographical coordinates and frequency of occurrence of the significant earthquakes dataset gathered by Lamontagne et al. (2018), a spatio-temporal point pattern dataset is created. We choose earthquakes that occurred after the year 1900 because older data are not complete and hence can highly affect the temporal component of the STPP. There are 137 significant earthquakes for the period between 1900-2017. We assume that the space on which the events are realized is continuous and the temporal component is discrete because earthquake events are identified annually. We follow a non-parametric approach to estimate the spatio-temporal intensity function, $\hat{\lambda}(\boldsymbol{x}, t)$, by using kernel functions, as suggested by Diggle (1985). We also follow a practical approach

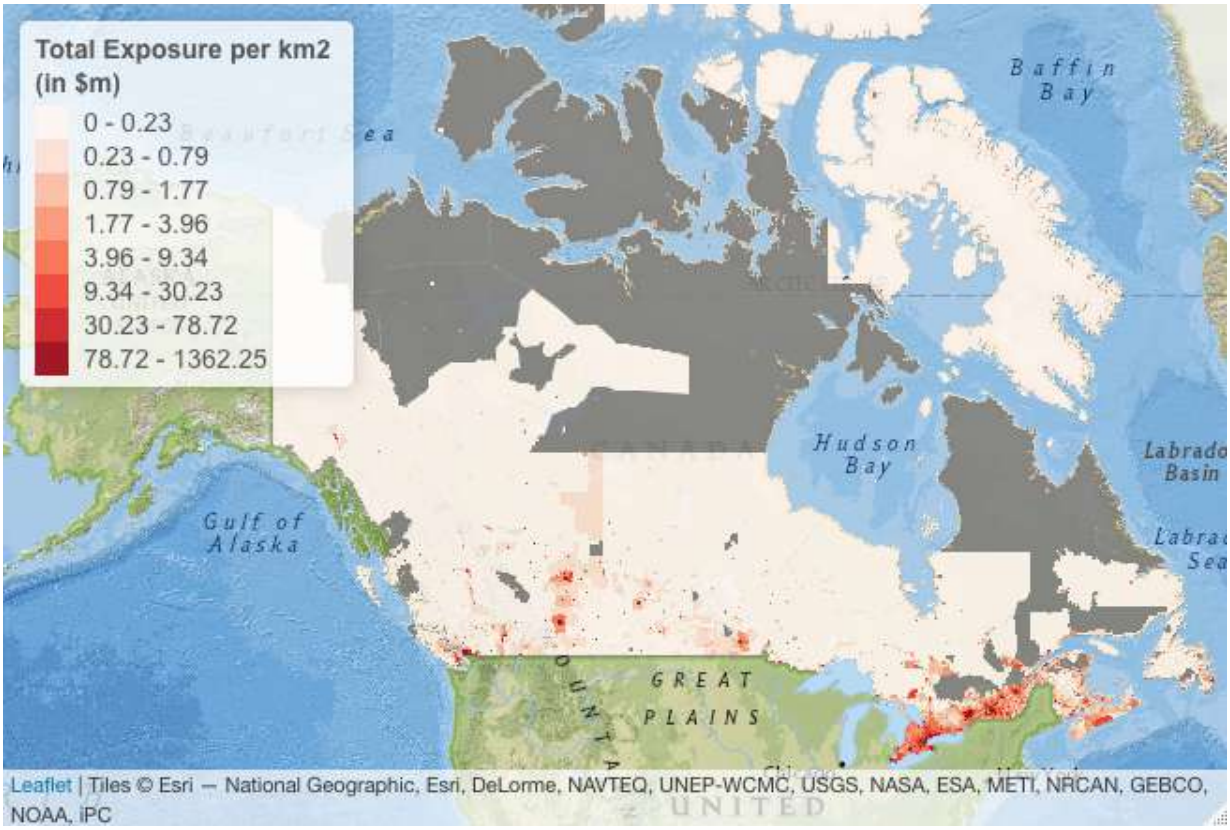


Figure 3.4: The calculated total exposure (building exposure and building contents exposure) of residential and non-residential buildings per km^2 for each CSD in Canada.

and assume that the STPP is separable [1] such that $\lambda(\boldsymbol{x}, t)$ can be factorized following

$$\lambda(\boldsymbol{x}, t) = \lambda_{\boldsymbol{X}}(\boldsymbol{x})\lambda_T(t), \tag{3.3.1}$$

for all $(\boldsymbol{x}, t) \in A \times T$. Given our interest in the $\text{PML}_{1/500}$, see eq. (3.1.1), which is a 1-in-500 year event, for earthquake damages, we are not interested in the exact time of occurrence of events, but rather the location of the epicenter and the affected communities. Hence, the separability assumption is valid for our project. Moreover, we will focus more on the spatial component, $\lambda_{\boldsymbol{X}}(\boldsymbol{x})$ and perform the proposed Voronoi residual analysis methods on it. Spatial modeling and simulations are performed using the `splancs`, `spatstat`, `deldir`, and `stpp` packages in **R** (Rowlingson and Diggle 2022; Baddeley and Turner 2005; Turner

1. The assumption is validated using Ghorbani et al. (2021)'s χ^2 test for first-order separability of spatio-temporal point patterns.

2021; Gabriel et al. 2022; R Core Team 2022), in addition to our functions for calculating the Voronoi residuals.

The temporal intensity function $\lambda_T(t)$ if estimated by using a Gaussian kernel with a bandwidth h_T that follows Silverman's rule-of-thumb (Silverman 1986), i.e.

$$h_T = 0.9An^{-1/5},$$

where $A = \min\{\text{sample standard deviation}, \text{sample interquartile range}/1.34\}$. As for the spatial intensity function $\lambda_{\boldsymbol{X}}(\boldsymbol{x})$, we choose the quartic kernel (Diggle 2013). The choice of the spatial bandwidth h determines the extent to which the intensity function will be smoothed, where larger bandwidths provide more smoothing. This choice involves a trade-off between bias and variance, such that larger bandwidths cause an increase in bias and a decrease in variance. The bias occurs because the estimator of the intensity function estimates a smoothed version of it rather than the intensity function itself. A large bandwidth requires a large number of points for estimation, thus reduces the estimation variance. The choice of a bandwidth is a controversial topic. We fit two different models for the spatial intensity function: both have the same kernel but with different bandwidths. The first bandwidth h_1 minimizes the estimated mean-square error of $\hat{\lambda}_{\boldsymbol{X}}(\boldsymbol{x})$. The second bandwidth h_2 is based on cross-validation. Define the cross-validation likelihood by

$$LCV(h_2) = \sum_l \log(\lambda_{-l}(\boldsymbol{x}_l)) - \int_A \lambda(\boldsymbol{u})\mathrm{d}\boldsymbol{u},$$

where $\lambda_{-l}(\boldsymbol{x}_l)$ is the leave-one-out kernel-smoothing estimate of the intensity at $\boldsymbol{x}_l$ with smoothing bandwidth h_2, $\lambda(\boldsymbol{u})$ is the kernel-smoothing estimate of the intensity at a spatial location $\boldsymbol{u}$ with smoothing bandwidth h_2 and A is the spatial window of observation. Baddeley, Rubak, and Turner (2015) observe that selecting a bandwidth by maximizing the LCV tends to choose more reasonable bandwidth values when the point pattern resembles

tight clusters, such as our earthquake data.

We assess the goodness-of-fit of the two fitted models, which we will call models H_1 and H_2, for the fitted quartic kernel with bandwidths h_1 and h_2, respectively. We also fit a homogeneous Poisson process, named model P, to the significant Canadian earthquakes spatial data set, where $\lambda(\boldsymbol{x}) = \lambda$ for all $\boldsymbol{x} \in A$. Model P will be used as a benchmark or scale to assess the Voronoi raw residuals plots of models H_1 and H_2.

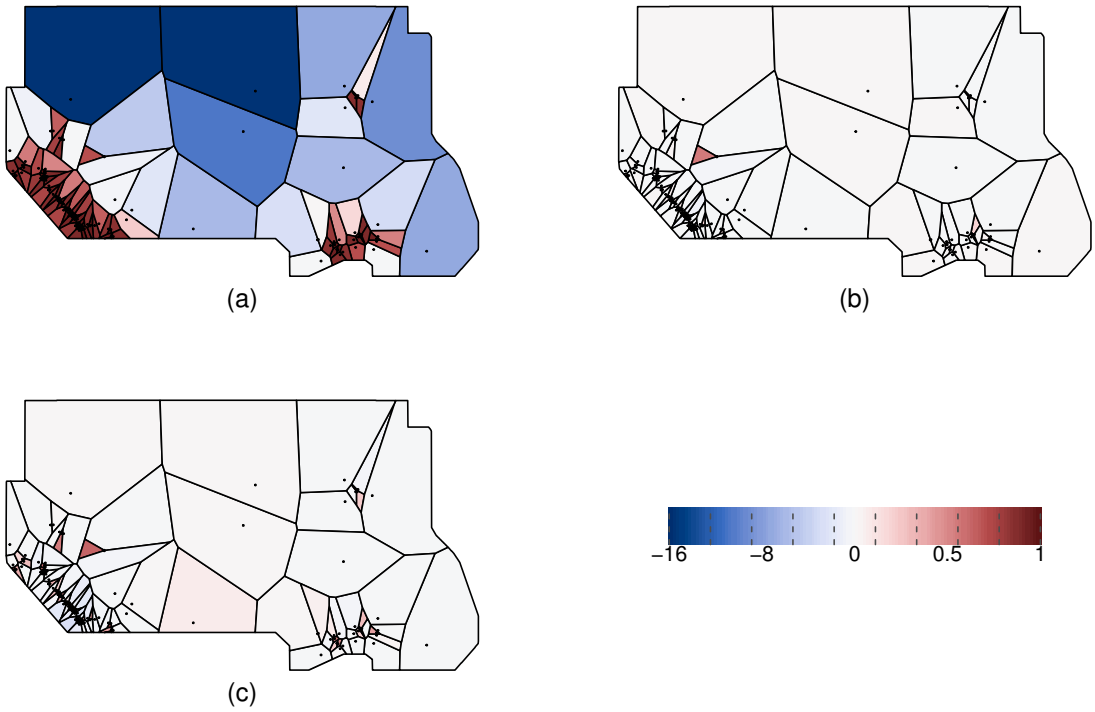


Figure 3.5: Raw Voronoi residuals of models P (a), H_1 (b), and H_2 (c).

Figure 3.5 shows the raw Voronoi residual plots on the same color scale for all fitted models. An accepted model is one that has raw Voronoi residuals close to 0. By definition, the raw Voronoi residuals cannot exceed 1, however, they can be negative as the model can overestimate. Model P is underestimating the expected count of earthquakes in the high-risk zones in Eastern and Western Canada, while it is overestimating the expected count in low-risk zones. We can expect a constant intensity function to perform poorly in

such data sets because the events are clustered in the known faults and they are rare in other areas. Based on the color scale for models H_1 and H_2, both outperform model P, in fact, the raw residuals are close to 0 in almost all polygons. However, the better of the two models is unclear. Focusing on the high-risk zones, the fit of model H_2 is slightly poorer than model H_1, as shown in Figures 6.1 and 6.2 in Appendix I.3. A similar conclusion is obtained from the visual inspection of Pearson Voronoi residual plots, see Figures 6.3, 6.4 and 6.5 in Appendix I.3.

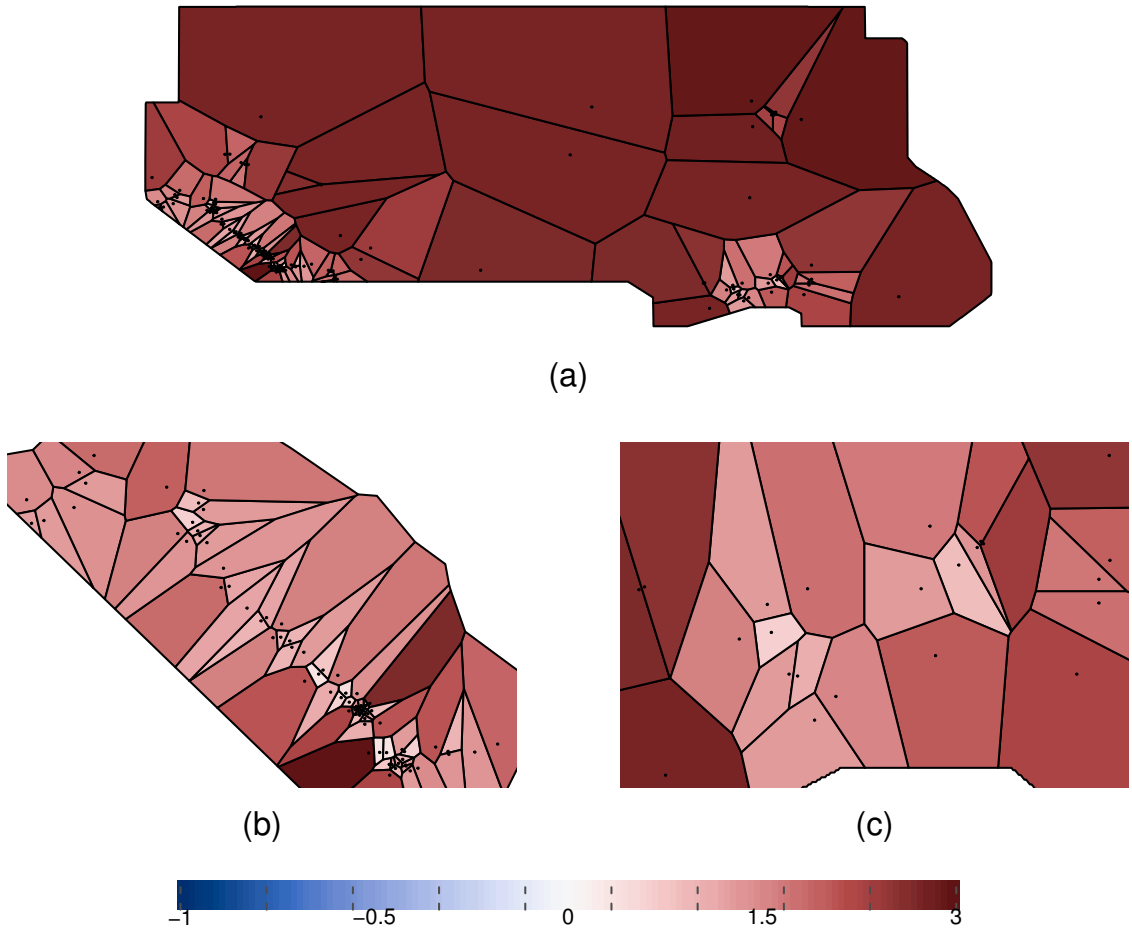


Figure 3.6: Deviance Voronoi residuals for model H_1 vs H_2 for Canada (a), Western Canada (b) and Eastern Canada (c).

To compare models H_1 and H_2, we use deviance Voronoi residuals, defined in eq. (3.2.1). Figure 3.6 shows the deviance Voronoi residuals for model H_1 versus model H_2. Voronoi

polygons with a positive (negative) deviance Voronoi residual indicate that model H_1 (H_2) provides a better fit. Inspecting the residual plot, we identify that model H_1 performs better than model H_2 in all polygons, and, in particular, model H_1 is performing much better than model H_2 in low-risk zones, such as Northern and Central Canada, where model H_2 tends to over-predict the frequency of occurrences. In both the Western and Eastern clusters of historical seismicity, more specifically, the Queen Charlotte fault (west) and the St. Lawrence paleo-rift faults (east), model H_1 is slightly outperforming model H_2. The polygon with the largest deviance residual is located in the Queen Charlotte fault, shown in Figure 3.6(b). There are no regions where model H_2 fits better than model H_1, in fact, the smallest difference in the log-likelihood is 0.341, which is also located in the Queen Charlotte fault. Overall, the log-likelihood ratio score for Canada is 210.67, implying a large improvement for model H_1 compared to model H_2. The log-likelihood ratio scores for Canada for model H_1 versus P is 586.56 and for model H_2 versus P is 375.89, indicating that both H_1 and H_2 models provide considerably more accurate estimations.

The estimated spatial intensity function $\hat{\lambda}_{\boldsymbol{X}}(\boldsymbol{x})$ for model H_1 and the temporal intensity $\hat{\lambda}_T(t)$ are multiplied to obtain the fitted spatio-temporal intensity function $\hat{\lambda}(\boldsymbol{x}, t)$, as given in eq. (3.3.1). We follow the simulation algorithm in the `stpp` package in **R** to simulate a large number of years of significant earthquakes. For more details on the spatio-temporal simulation algorithm, see Gabriel et al. (2022).

(iii) Estimation of ground shaking intensity

Earthquake hazard across Canada is evaluated by seismologists and geophysicists and the nation's seismic hazard maps are updated regularly. The National Building Code is subsequently revised to minimize the seismic risk on new infrastructures. Calculation of seismic risk requires the estimation of the ground shaking intensity, which can be defined by attenuation relationships of a given ground motion index, such as

the peak ground acceleration (PGA), peak ground velocity, or spectral acceleration. The Geological Survey of Canada (GSC) provides seismic hazard values for a grid extending over Canada and surrounding areas for the 2015 National Building Code of Canada (NBCC) (Halchuk, Adams, and Allen 2015). We use the mean PGA, which is available on firm soil sites, evaluated at the following probabilities of exceedance: $0.02, 0.01375, 0.0100, 0.00445, 0.0021, 0.0010, 0.0005$ and 0.000404 per annum. For each grid point, a Generalized Pareto Distribution (GPD) is fitted such that the given quantiles represent the return levels that are exceeded every $50-2475$ years. The quantile of a GPD is shown in eq. (I.1) in Appendix I.4.

For a simulated earthquake location as in (ii), the GPD parameters of the grid point that is nearest to the earthquake epicenter are used to simulate a PGA value. To match the definition of significant earthquakes in Canada (Lamontagne et al. 2018), the PGA value must have an associated moment magnitude $\mathrm{M} > 6$. The corresponding MMI level is calculated by using the relationship $\mathrm{MMI} = 3.66\log(\mathrm{PGA}) - 1.66$, with standard error 1.08, for PGA in $\mathrm{cm/s^2}$ (Wald et al. 1999). The corresponding moment magnitude M is calculated following Bakun's predicted distance attenuation equations,

- Eastern Canada (Bakun, Johnston, and Hopper 2003):

$$\mathrm{M} = \frac{1}{1.68}\left(\mathrm{MMI} - 1.41 + 0.00345\mathrm{d} + 2.08\log_{10}(\mathrm{d})\right), \tag{3.3.2}$$

- Western Canada (Bakun and Wentworth 1997):

$$\mathrm{M} = \frac{1}{1.09}\left(\mathrm{MMI} - 5.07 + 3.69\log_{10}(\mathrm{d})\right), \tag{3.3.3}$$

where d is the distance in kilometers from the epicenter of the earthquake to its nearest neighbor on the PGA grid. Even though an earthquake has a single magnitude, it has multiple intensity values, depending on the location from the epicenter. Accordingly,

Bakun's predicted distance attenuation equations are used to produce isoseismal maps, which are maps that identify areas of equally felt seismic intensity. These are estimated by calculating the distance from the epicenter for each MMI level, based on eq. (3.3.2) and eq. (3.3.3), and MMI circles with radii equivalent to the calculated distances are graphed accordingly. The MMI levels are very important information that are required to quantify the percentage of damage caused by an earthquake. We consider the uncertainty in the estimation by applying random noise with the respective standard errors, obtained from Bakun and Wentworth (1997), Bakun, Johnston, and Hopper (2003), and Wald et al. (1999). The CSDs inside each MMI circle are identified and the percentage of their affected areas are calculated for each seismic intensity level.

(iv) Calculating damage rates

Earthquake damage can result in financial losses, such as the cost of damage repair and losses due to business interruptions, in addition to non-financial losses, such as fatalities and injuries. Each category of losses is divided further into direct losses from the damage caused by the ground shake and indirect losses from damage due to other hazards induced by the earthquake, such as a tsunami, landslide, liquefaction and/or a fire. This distinction is of interest to insurers to process claims depending on the coverage and policy conditions. In this article, only the direct financial losses that occur due to earthquake damage to buildings are considered. Building damage is split into structural damage and non-structural damage. Structural damage is the damage to the skeleton of the building, such as the roof and load-bearing walls. Non-structural damage can affect drift-sensitive components, such as non-bearing walls/partitions, veneer and finishes, acceleration-sensitive components, such as piping systems, elevators and lightning fixtures, and finally building contents, such as bookcases, office equipment and furnishings.

The Applied Technology Council, ATC (1985), provides a benchmark study that relies on

MMI to quantify the degree of ground shaking. Motion-damage relationships are developed, where the probability of being in a defined damage state for different levels of MMI is calculated. To quantify the damage due to an earthquake, MMI-based damage probability matrices (DPMs) for structural and non-structural damage for each building classification are required. These matrices provide for each building class the probability that a building is in one of seven damage states, given the MMI at the location of the building. The damage states and their corresponding damage factor range, which is the range of the percentage of exposure value that is damaged, are shown in the first two columns of Table 3.1. Thibert (2008) developed DPMs for structural and non-structural damage in BC by assuming that the buildings are nearly regular in shape, founded on firm ground, and are designed based on a National Building Code prior to 1990. Table 3.1 provides an example of a DPM.

Table 3.1: DPM for Structural Damage in Wood Light Frame Residential building (Thibert 2008)

Damage state	**Damage factor range**	**VI**	**VII**	**VIII**	**IX**	**X**	**XI**	**XII**
No damage	0	0.08	0.04	0.01	***	***	***	***
Slight damage	$0-1$	0.75	0.28	0.06	0.01	***	***	***
Light	$1-10$	0.17	0.64	0.86	0.69	0.19	0.02	***
Moderate	$10-30$	***	0.04	0.05	0.20	0.76	0.69	0.42
Heavy	$30-60$	***	***	0.02	0.10	0.12	0.25	0.50
Major	$60-100$	***	***	***	***	0.02	0.04	0.06
Destroyed	100	***	***	***	***	***	***	0.02

*** represents very small probability values, almost 0.

The percentage of damage for each building classification is defined in terms of the mean damage factor (MDF), which quantifies the expected damage as a percentage of exposure value. The MDF is the expected value of the damage given an MMI level, and is calculated by adding up the product of the damage factor by its corresponding probability. To embed further variability in the simulations, the damage factor is randomly drawn from each damage range. For each CSD in an MMI circle, the MDF for each damage type (structural (S), acceleration-sensitive (AS) non-structural, drift-sensitive (DS) non-structural, and

building contents (BldgC)) is estimated for each building class.

(v) Estimating seismic losses and insurance claim payments

The losses are calculated by summing the losses generated by S, DS, AS, and BldgC damages. Following Onur, Ventura, and Finn (2005), we split the building exposure over damage types following: 25% of the building exposure is for S, 37.5% is for DS and 37.5% is for AS components. Accordingly, the losses for the j^{th} CSD are simply the product of the MDF for a given MMI level and the exposure of the CSD at that MMI level such that

$$\begin{aligned}
\mathrm{L}_{j,k,S} &= \left[\mathrm{MDF}_{S,k} \times \left(0.25 \times \text{building exposure at } \mathrm{CSD}_{j,k}\right)\right], \\
\mathrm{L}_{j,k,DS} &= \left[\mathrm{MDF}_{DS,k} \times \left(0.375 \times \text{building exposure at } \mathrm{CSD}_{j,k}\right)\right], \\
\mathrm{L}_{j,k,AS} &= \left[\mathrm{MDF}_{AS,k} \times \left(0.375 \times \text{building exposure at } \mathrm{CSD}_{j,k}\right)\right], \\
\mathrm{L}_{j,k,BldgC} &= \left[\mathrm{MDF}_{BldgC,k} \times \text{building contents exposure at } \mathrm{CSD}_{j,k}\right],
\end{aligned}$$

where k is an index for the k^{th} building class. Thus,

$$\text{Total Losses}_{j,k} = \mathrm{L}_{j,k,S} + \mathrm{L}_{j,k,DS} + \mathrm{L}_{j,k,AS} + \mathrm{L}_{j,k,BldgC}.$$

Note that a CSD can be affected by multiple MMI levels, and accordingly the percentage area affected at each MMI level is used to distribute the exposure of a CSD.

The insurance terms used in this article are based on the deductibles, policy limits and insurance market penetration rates provided in AIR Worldwide (2013), which are displayed in Table 3.2. AIR Worldwide have only given statistics for QC and BC, and hence, we assume the smallest market penetration for the provinces for which we have no information.

Insurance claims for the j^{th} CSD are calculated following

$$\text{ClaimPmt}_{j,k} = \begin{cases} 0, & \text{Total Losses}_{j,k} \leq d_{j,k}, \\ \pi_j \left(\text{Total Losses}_{j,k} - d_{j,k}\right), & d_{j,k} < \text{Total Losses}_{j,k} \leq u_{j,k}, \\ \pi_j \left(u_{j,k} - d_{j,k}\right), & \text{Total Losses}_{j,k} > u_{j,k}, \end{cases}$$

where

$$\begin{aligned} \pi_j &= \%\text{ earthquake insurance market penetration at CSD}_j, \\ d_{j,k} &= \%\text{ deductible}_j \times \sum_k \left(\text{total exposure at CSD}_{j,k}\right), \\ u_{j,k} &= \%\text{ policy limit}_j \times \sum_k \left(\text{total exposure at CSD}_{j,k}\right), \end{aligned}$$

where k is an index for the k^{th} building class and total exposure for a CSD includes building content exposure. Finally,

$$\begin{aligned} \text{Total Losses}_j &= \sum_k \text{Total Losses}_{j,k}, \\ \text{ClaimPmt}_j &= \sum_k \text{ClaimPmt}_{j,k}. \end{aligned}$$

Appendix I.5 summarizes the methodology in two algorithms: Algorithm 1 includes the steps required to simulate an earthquake and Algorithm 2 explains how to calculate the losses and the insurance claims for each simulated earthquake.

3.3.2 MCT for earthquake risk

OSFI's MCT formula in eq. (3.1.1) is a function of the PML in Eastern and Western Canada. Actuaries define the PML as the worst case scenario of the losses. Let $X_1, \ldots, X_n$ be a sequence of independent random variables having a common distribution function

Table 3.2: Assumptions for the earthquake insurance terms and market penetration rates, as prescribed in AIR Worldwide (2013)

Property type	**Location**	**Market penetration rate**	**Deductible**	**Limit**
Residential	Vancouver Metro	55%	10%	100%
	Victoria Metro	70%	8%	100%
	Rest of BC	40%	8%	100%
	Montréal Metro	5%	5%	100%
	Québec Metro	2%	5%	100%
	Rest of QC	2%	5%	100%
Commercial / Industrial	Vancouver Metro	85%	10%	80%
	Victoria Metro	85%	7.5%	80%
	Rest of BC	85%	7.5%	80%
	Montréal Metro	60%	5%	80%
	Québec Metro	60%	5%	80%
	Rest of QC	60%	5%	80%

Market penetration rates are percentages of the total exposure value.
Deductibles and policy limits are percentages of the insured exposure value.

F and consider $M_n = \max\{X_1, \ldots, X_n\}$. There are several suggestions in the actuarial literature to calculate the PML, such as $(1+\theta)\mathbb{E}[M_n]$ or $\mathbb{E}[M_n] + \theta\sqrt{\text{Var}[M_n]}$, as suggested by Wilkinson (1982) and Kremer (1990, 1994), where θ is a chosen constant. The PML is mathematically defined to be an extreme quantile of M_n such that

$$\mathbb{P}(M_n \leq \text{PML}_\epsilon) = 1 - \epsilon \tag{3.3.4}$$

for some small $\epsilon > 0$. Accordingly, the PML is the threshold dollar value of losses beyond which losses are highly unlikely.

As explained in Appendix I.4, under some conditions, the distribution of the exceedances over a high threshold u can be approximated by a GPD $G_{\xi,\sigma}$ and the number N of exceedances over u follows a Poisson distribution with rate λ. The distribution of the maximum of the N exceedances can be approximated by a Generalized Extreme Value (GEV) distribution $H_{\xi,\mu,\psi}$ with $\mu = \frac{\sigma}{\xi}\left(\lambda^\xi - 1\right)$ and $\psi = \sigma\lambda^\xi$ (Cebrián, Denuit, and Lambert

2003). Solving for the PML in eq. (3.3.4) yields

$$\mathrm{PML}_\epsilon = u + \frac{\sigma}{\xi}\left[\left(-\frac{\lambda}{\ln(1-\epsilon)}\right)^{\xi} - 1\right]. \tag{3.3.5}$$

See Section I.4 in Appendix I.4 for details.

Inspired by Solvency II, which guides the European standards for capital requirements, we propose using the correlation of the insured losses between Canadian provinces and territories to calculate the minimum capital requirements for earthquake risk in Canada. The correlation coefficients should reflect the dependence between earthquake financial risks in Canadian provinces and territories. The Solvency II capital requirements for earthquake insurance risk adapted to Canada is

$$\text{Country-wide}\,\mathrm{PML}_{1/500} = \sqrt{\sum_{r,s} \mathrm{CorrEQ}_{r,s} \times \mathrm{PML}_{1/500,r} \times \mathrm{PML}_{1/500,s}}, \tag{3.3.6}$$

where:

1. the sum includes all possible combinations (r, s) of Canadian provinces and territories.
2. $\mathrm{CorrEQ}_{r,s}$ denotes the correlation of the insured losses for earthquake risk for province r and province s
3. $\mathrm{PML}_{1/500,r}$ and $\mathrm{PML}_{1/500,s}$ denote the Gross PML for province r and s respectively, which is the PML amount after deductibles but before any reinsurance deductions.

The $\mathrm{PML}_{1/500}$ is calculated both empirically from the simulations by computing the $(1 - 1/500)$ quantile of the annual maximum losses, which may be zero for years with no earthquake damage and by substituting estimates of the parameters μ, σ, ξ, and λ in eq. (3.3.5). The correlation matrix CorrEQ is created by calculating the pairwise correlation of the simulated financial losses. We should have high correlation when

provinces are affected concurrently with the same earthquake. Several correlation measures were attempted: Pearson's correlation coefficient, Kendall's tau and Kendall's tau for zero-inflated continuous variables (Pimentel, Niewiadomska-Bugaj, and Wang 2015). The latter was considered because of the nature of the data where a large number of observations have no financial losses because earthquakes do not affect all provinces in Canada simultaneously[2].

3.4 Results

In this section we compare our results from the proposed methodology in Section 3.3 to other sources and articles. Section 3.4.1 compares the building exposure values following our methodology to the exposure values obtained from CatIQ, Canada's insured loss and exposure indices provider. Section 3.4.2 compares the financial losses and insurance claims generated from our algorithm to another earthquake study in the literature; see Onur, Ventura, and Finn (2005). Finally, Section 3.4.3 provides a detailed analysis of the earthquake simulation results.

3.4.1 Comparison of exposure values

CatIQ collects values of exposure, insured losses and other related information to serve the needs of insurers, reinsurers and other stakeholders through an online subscription-based platform. Figure 3.7 compares the exposure values obtained from CatIQ for 2020 to the insured exposure calculated by applying the earthquake insurance market penetration rates from Table 3.2 to the cost of building replacement, as explained in Section 3.3.1. The comparison is made in June 2021 dollar value. The values on the y-axis are removed to

2. As discussed in Denuit and Mesfioui (2017), there are bounds on Kendall's tau for zero-inflated continuous variables such that the correlation values are not between $[-1, 1]$, but rather have a smaller range. Accordingly smaller Canada-wide PML values will be obtained, compared to the other methods. Hence, this correlation method was ultimately deemed inappropriate for the purpose of this article.

maintain the confidentiality of CatIQ's data, which represents aggregated exposure values from a majority of Canadian insurers. Our methodology is based on estimations and several assumptions, while CatIQ's exposure values are an accurate representation of the insured exposure for the insurers who submitted their data. Given that we intend to provide an open-source earthquake insurance risk assessment tool that can be easily explained and there are no publicly available exposure data, the insured exposure values for residential buildings calculated with our methodology are deemed sufficient for this paper. The non-residential exposure differs considerably, which is explained by the use of many assumptions to estimate the values; see Section 3.3.1.

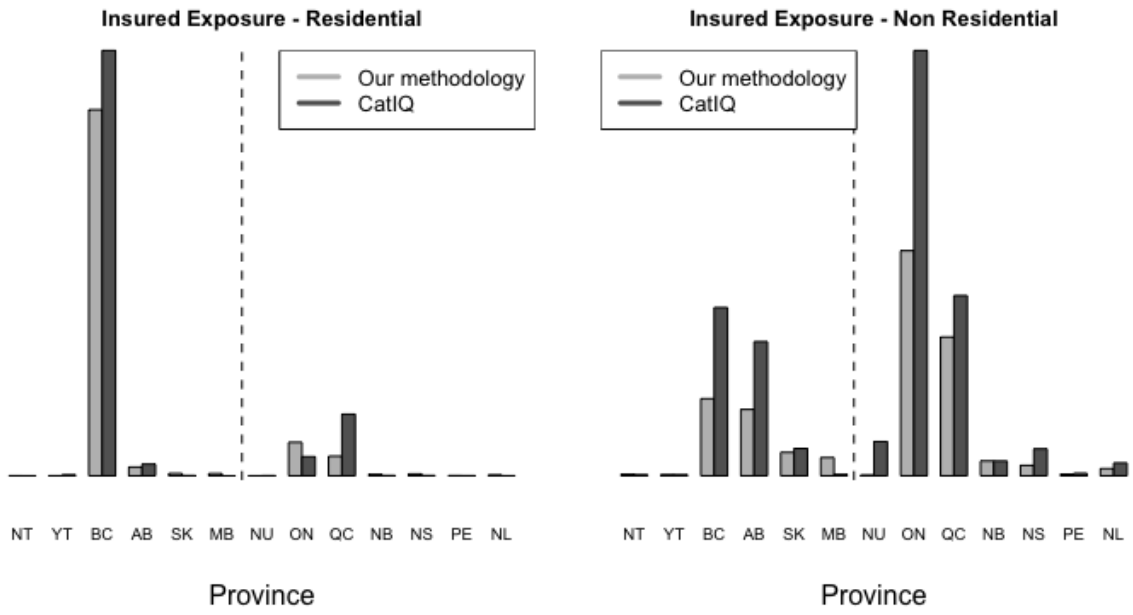


Figure 3.7: Calculated insured exposure, as explained in Section 3.3.1, vs the exposure collected by CatIQ.

3.4.2 Comparison of the simulated financial losses with other earthquake studies

In this section, we compare the results of the methodology explained in Section 3.3 to calculate direct financial losses and insurance claims with those estimated in earthquake

studies in the literature.

Conducted in 2001 and published in 2005, the study by Onur, Ventura, and Finn (2005) estimates the financial losses for the City of Victoria and the City of Vancouver, subject to an earthquake of MMI VIII affecting the entirety of the cities. Total monetary losses of $430 million for the City of Victoria and $3.5 billion for the City of Vancouver, in 2001 Canadian dollars, are estimated. Inflated to June 2021 by using the BCPI from Statistics Canada, Government of Canada (2020b), the values correspond to $835 million for the City of Victoria and $6.8 billion for the City of Vancouver. Our methodology, under the same MMI level assumption, produces losses of $941 million for the City of Victoria and $5.73 billion for the City of Vancouver. The differences can be attributed to two factors: the losses calculated by Onur, Ventura, and Finn (2005) excluded the losses resulting from damage to building contents, and different sources in calculating the construction costs where Onur, Ventura, and Finn (2005) used values provided by local construction companies.

3.4.3 Simulation results

In this section, we give the results of our earthquake simulations. We also compare the value of the Canada-wide PML for earthquake risk in P&C companies obtained from our proposed method in eq. (3.3.6) to that computed from OSFI's formula in eq. (3.1.1).

100,000 years of earthquake simulations are performed by relying on the procedure detailed in Section 3.3. The number of years is chosen to be high enough to minimize the variance in the results[3]. Table 3.3 summarizes the proportion of years that had a significant earthquake and the proportion of years that had a damage-inducing significant earthquake.

Figure 3.8 shows the locations and moment magnitudes of a sample of 200 years from the

3. The results in this section are compared against another set of 100,000 years and minimal variations were observed.

Table 3.3: The proportion of years with a significant earthquake and the proportion of years with a significant earthquake causing damage, based on 100,000 simulated years of earthquakes.

Number of earthquakes (n)	**% of years with n earthquakes**	**% of years with n earthquakes causing damage**
0	45.651	59.536
1	27.127	26.321
2	15.265	9.886
3	7.127	3.117
4	2.999	0.846
5	1.177	0.227
6	0.425	0.057
7	0.160	0.01
8	0.065	0
9	0.013	0
10	0.003	0
11	0.001	0
12+	0	0

simulations. There is a great similarity with the historical seismicity shown in Figure 3.1 in terms of the locations and intensity of the earthquakes.

Following the methodology explained in Section 3.3.1 to calculate exposure and Algorithms 1 and 2, we obtain values for financial losses and insurance claims for each simulated earthquake. Figure 3.9 illustrates the expected value of the size of the simulated financial losses and insurance claims for each CSD, conditioning on the occurrence of an earthquake, whereas Figures 3.10 and 3.11 compare the expected value per province. Some CSDs are located in zones free of any seismic activities such that they do not witness any significant earthquakes in the 100,000 simulated years and accordingly they observe no damage. The expected value of financial losses, conditional on the occurrence of an earthquake, for each province is relatively proportionate and consistent with the exposure values provided in Figure 3.3. Locations with high value of exposure witness high values of financial losses, however, insurance claims are affected by the insurance market penetration and the insurance terms. Given the large proportion of earthquake insurance market penetration for

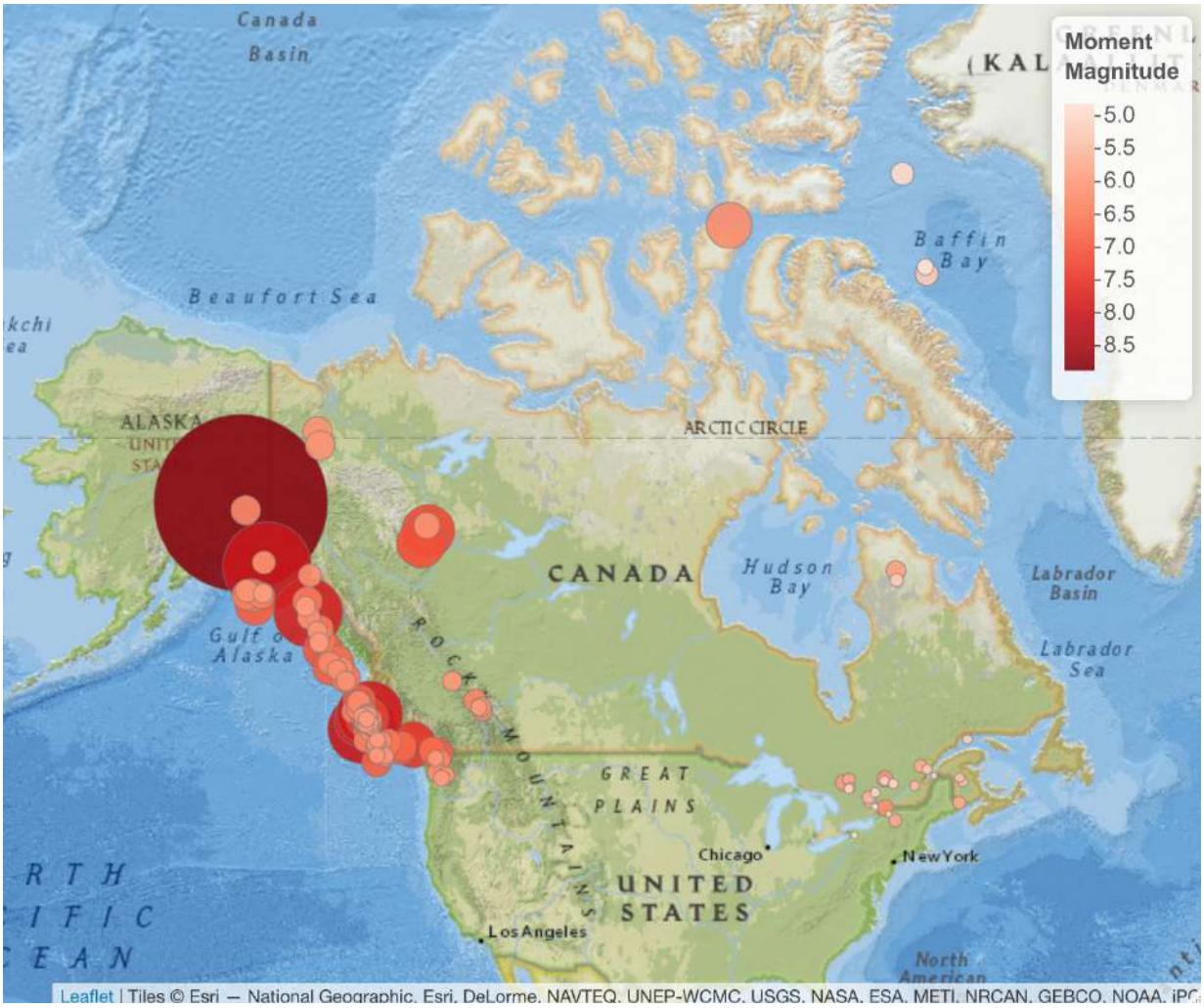
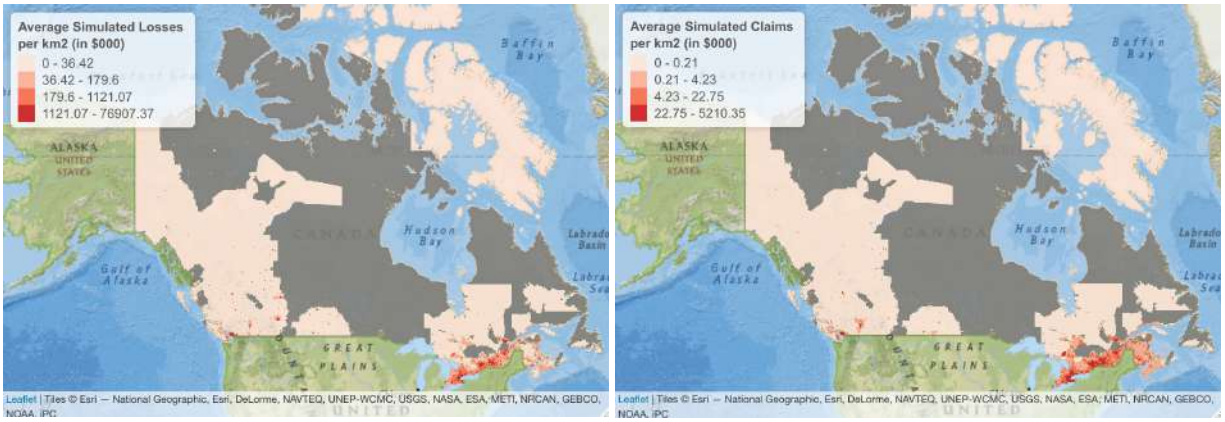


Figure 3.8: 200 years of simulated earthquakes. The size and color of the circles are proportionate to the moment magnitude.

non-residential buildings in Eastern Canada compared to residential buildings, we observe a surge in the expected value of insurance claims for non-residential buildings. Larger losses are seen in Eastern Canada due to the predicted distance attenuation formulas, presented in eq. (3.3.2) and eq. (3.3.3). For an earthquake of moment magnitude 6, MMI VI can be reached at a distance of 200 km from the epicenter in Eastern Canada, compared to a distance of 33 km from the epicenter in Western Canada. As discussed in U.S. Geological Survey (2018), seismic energy travels in Eastern North America much further than in Western North America.



(a) Average Simulated Losses (b) Average Simulated Claims

Figure 3.9: Average financial losses and insurance claims, conditional on the occurrence of an earthquake, based on simulated 100,000 years.

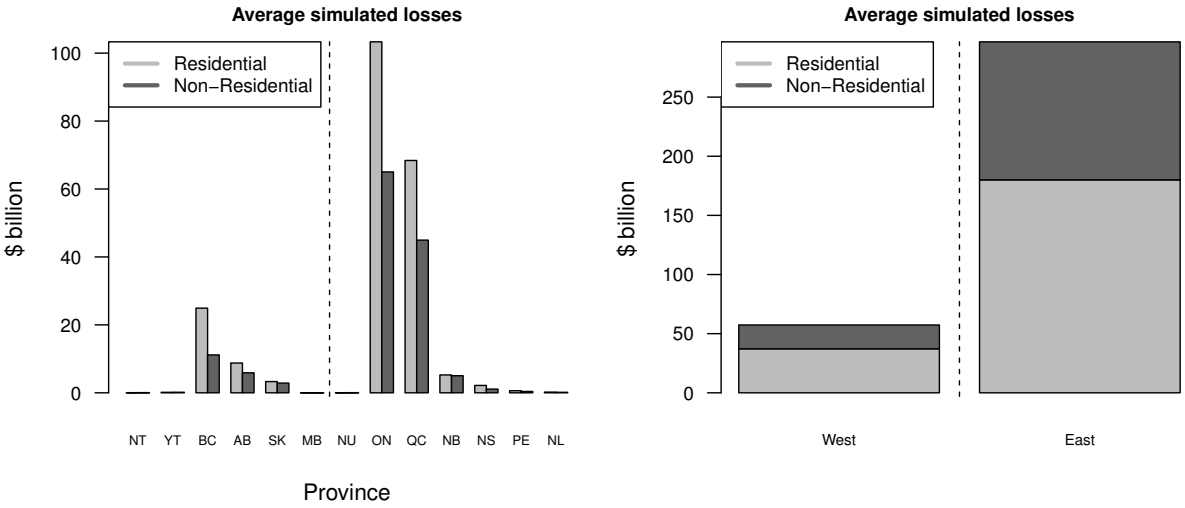


Figure 3.10: Average financial losses per province, conditional on the occurrence of an earthquake, based on 100,000 simulated years. The vertical dashed line splits Eastern and Western provinces.

In addition to the PML for the insurance claims, we estimate the Canada-wide PML for financial losses, which include the insured and non-insured portions of the losses, to provide a big picture of the damage. The correlation of the financial losses and insurance claims between the provinces is required for eq. (3.3.6). Table 3.4 and Tables 6.3, 6.4 and 6.5 in Appendix I.6 provide Pearson's correlation coefficients of the financial losses, Pearson's correlation coefficients of the insurance claims, Kendall's tau of the financial losses and

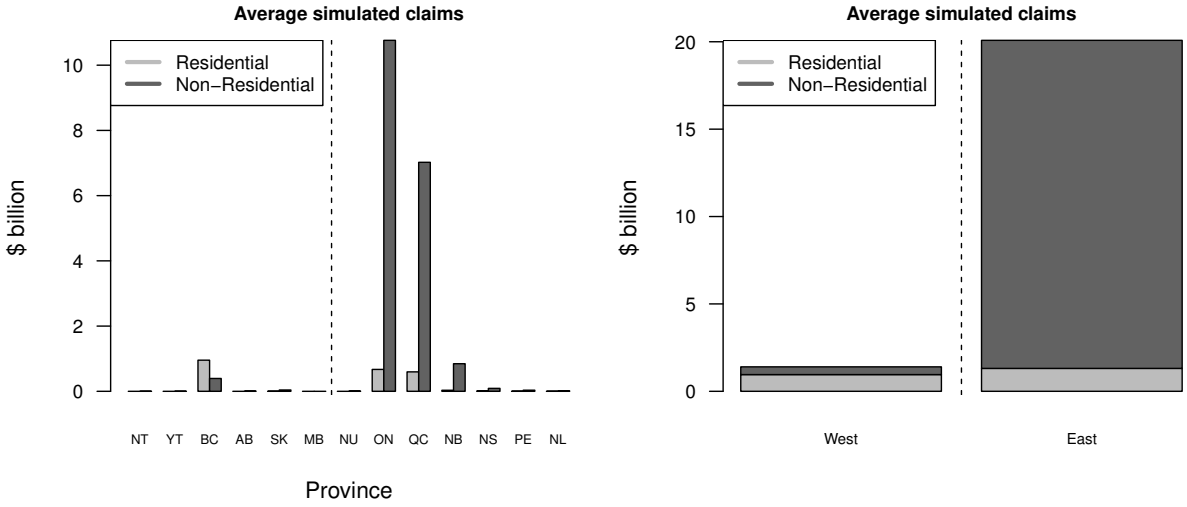


Figure 3.11: Average insurance claims per province, conditional on the occurrence of an earthquake, based on 100,000 simulated years. The vertical dashed line splits Eastern and Western provinces.

Kendall's tau of the insurance claims, respectively.

Table 3.4: Pearson correlation of the simulated financial losses between Canadian provinces, based on 100,000 years of simulated earthquakes.

	NL	**PE**	**NS**	**NB**	**QC**	**ON**	**MB**	**SK**	**BC**	**YT**	**NT**	**AB**	**NU**
NL	1.00	0.38	0.32	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PE	0.38	1.00	0.81	0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NS	0.32	0.81	1.00	0.82	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NB	0.31	0.91	0.82	1.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
QC	0.00	0.00	0.00	0.03	1.00	0.69	0.64	0.00	0.00	0.00	0.00	0.00	0.00
ON	0.00	0.00	0.00	0.00	0.69	1.00	0.64	0.00	0.00	0.00	0.00	0.00	0.00
MB	0.00	0.00	0.00	0.00	0.64	0.64	1.00	0.03	0.00	0.03	0.02	0.00	0.02
SK	0.00	0.00	0.00	0.00	0.00	0.00	0.03	1.00	0.04	0.02	0.02	0.08	0.02
BC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	1.00	0.66	0.53	0.80	0.55
YT	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.02	0.66	1.00	0.87	0.40	0.88
NT	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.53	0.87	1.00	0.32	0.97
AB	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.80	0.40	0.32	1.00	0.35
NU	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.55	0.88	0.97	0.35	1.00

Table 3.5 shows the parameter estimates of the fitted marked homogeneous Poisson process with rate λ, shape parameter ξ and scale parameter σ for the size of the excesses over a high threshold u. The parameters and their standard errors (s.e.) are estimated from the

simulated financial losses and insurance claims for each Canadian province, for Eastern and Western Canada and for the whole country. The thresholds are chosen to be the 0.95 quantile of the simulated data, or the 0.9 quantile for provinces with a small number of exceedances. Diagnostic plots (not shown) are used to confirm the adequacy of the fitted GPD models.

Tables 3.6 and 3.7 summarize the $\text{PML}_{1/x}$ for $x \in \{100, 250, 500, 750, 1000\}$ of the direct financial losses and the insurance claims, respectively. Results are provided for each province and for Eastern and Western Canada. The $\text{PML}_{1/x}$ are computed following two methods: by calculating the appropriate quantiles from the simulated data for each province (simulated), and by plugging in the parameter estimates from Table 3.5 in eq. (3.3.5) (estimated). Additionally, Canada-wide $\text{PML}_{1/x}$ is calculated in three ways: following OSFI's formula in eq. (3.1.1), and the proposed formula in eq. (3.3.6) calculated with Pearson's correlation coefficient and Kendall's tau. We observe that the simulated and the estimated results are very similar, confirming the adequacy of the fitted models. We also observe that the proposed MCT formula in eq. (3.3.6) produces values that are comparable to OSFI's current approach in eq. (3.1.1) and that results using Kendall's tau are more conservative than results using Pearson correlation. This is be explained by the larger values for Kendall's tau compared to Pearson Correlation in Eastern Canadian provinces, more specifically for QC.

Figure 3.12 compares the Canada-wide $\text{PML}_{1/x}$ for $x \in [100, 1000]$ by using eq. (3.1.1) and eq. (3.3.6) with Pearson correlation and with Kendall's tau. We observe that for the financial losses, the proposed formula is more conservative than OSFI's for any value of x, where the difference gap increases as x increases, with Kendall's tau producing higher values than Pearson's correlation coefficient. For $\text{PML}_{1/500}$, eq. (3.3.6) with Pearson correlation is \$27 billion more than OSFI's formula, compared to a \$51 billion difference when Kendall's tau is used. This is explained by the fact that the proposed method

captures the strong dependence between QC and ON, where both provinces are affected concurrently by the same earthquakes due to the distance attenuation equations in Eastern Canada producing larger isoseismal maps than Western Canada. However, for the $\text{PML}_{1/x}$ of insurance claims, eq. (3.3.6) with Pearson correlation coefficient produces very similar results to OSFI's, while Kendall's tau offers slightly higher values. In fact, for $\text{PML}_{1/500}$, eq. (3.3.6) with Kendall's tau is higher by approximately $3 billion. Should Canadian P&C insurance companies implement the proposed equation, especially with Pearson correlation coefficient, the transition will be smooth due to the resemblance in the values between both methods.

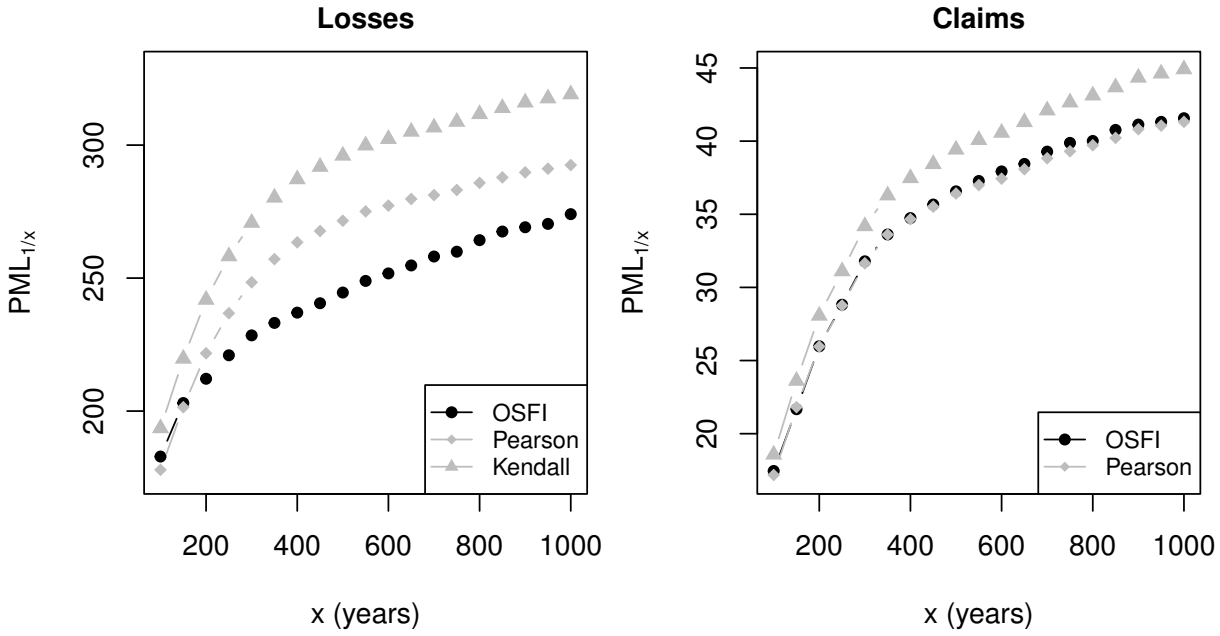


Figure 3.12: Canada-wide $\text{PML}_{1/x}$: OSFI vs eq. (3.3.6) with Pearson's correlation coefficient and Kendall's tau

Table 3.5: Estimates of the parameters of the fitted homogeneous Poisson process for the simulated financial losses and insurance claims, based on 100,000 simulated years.

Province	**Financial losses**			**Insurance claims**		
	σ (s.e.)	ξ (s.e.)	λ (s.e.)	σ (s.e.)	ξ (s.e.)	λ (s.e.)
NL	0.0836 (0.0067)	0.0032 (0.0618)	0.0040 (0.0002)	0.0052 (0.0005)	0.3388 (0.0890)	0.0030 (0.0002)
PE	0.1832 (0.0197)	-0.0637 (0.0794)	0.0019 (0.0001)	0.0254 (0.0025)	0.2751 (0.0836)	0.0036 (0.0002)
NS	1.5970 (0.1376)	-0.2589 (0.0556)	0.0022 (0.0001)	0.1369 (0.0133)	0.0109 (0.0667)	0.0020 (0.0001)
NB	2.3892 (0.1197)	-0.1398 (0.0337)	0.0072 (0.0003)	0.7719 (0.0424)	-0.0566 (0.0399)	0.0070 (0.0003)
QC	31.5666 (1.2992)	-0.1804 (0.0262)	0.0096 (0.0003)	11.1067 (0.4390)	-0.2301 (0.0238)	0.0095 (0.0003)
ON	38.4570 (1.4825)	-0.1784 (0.0197)	0.0088 (0.0003)	2.4030 (0.1313)	0.2751 (0.0437)	0.0086 (0.0003)
MB	0.4009 (0.0166)	-0.3343 (0.0200)	0.0067 (0.0003)	0.1173 (0.0051)	-0.3647 (0.0208)	0.0051 (0.0002)
SK	0.0801 (0.0107)	1.6201 (0.1526)	0.0030 (0.0002)	0.0042 (0.0005)	0.8859 (0.1369)	0.0017 (0.0001)
BC	2.8283 (0.0800)	-0.1021 (0.0174)	0.0200 (0.0004)	0.1170 (0.0056)	0.6930 (0.0444)	0.0157 (0.0004)
YT	0.9433 (0.0620)	-0.1498 (0.0434)	0.0040 (0.0002)	0.0903 (0.0118)	0.6248 (0.1217)	0.0026 (0.0002)
NT	6.8276 (0.5989)	-0.0433 (0.0609)	0.0025 (0.0002)	1.4643 (0.1392)	0.2565 (0.0770)	0.0030 (0.0002)
AB	0.2505 (0.0207)	0.2880 (0.0638)	0.0033 (0.0002)	0.1236 (0.0128)	-0.0426 (0.0764)	0.0020 (0.0001)
NU	2.0897 (0.1620)	-0.1284 (0.0458)	0.0025 (0.0002)	0.0988 (0.0088)	0.7871 (0.0841)	0.0049 (0.0002)
East	38.9950 (1.4906)	-0.0827 (0.0222)	0.0103 (0.0003)	11.9035 (0.4675)	-0.1046 (0.0238)	0.0101 (0.0003)
West	11.8276 (0.4373)	0.0851 (0.0309)	0.0243 (0.0005)	0.1912 (0.0088)	0.9903 (0.0462)	0.0198 (0.0004)
Total	39.5434 (1.3060)	-0.0752 (0.0206)	0.0149 (0.0004)	12.6883 (0.5151)	-0.1385 (0.0194)	0.0077 (0.0003)

Table 3.6: $\text{PML}_{1/x}$ values (in $ billions) for the financial losses, based on 100,000 simulated years.

	x	NL	PE	NS	NB	QC	ON	MB	SK	BC	YT	NT	AB	NU	East	West	OSFI	Pearson	Kendall
Simulated	100	0.1	0.9	2.6	10.4	135.0	53.4	0.8	0.0	4.0	0.9	9.7	0.4	1.3	180.1	14.9	182.9	177.9	193.6
	250	0.1	1.4	4.3	12.6	164.6	87.3	1.1	0.1	6.2	1.7	19.5	0.7	3.8	214.2	28.1	221.0	236.8	258.2
	500	0.2	1.6	5.6	13.6	180.5	108.6	1.2	0.1	7.7	2.4	25.7	0.9	5.4	234.4	38.1	244.6	271.6	296.0
	750	0.2	1.7	6.2	14.2	185.7	115.4	1.3	0.2	8.5	2.7	28.5	1.0	6.3	248.4	42.3	259.9	283.1	308.8
	1000	0.3	1.7	6.6	14.7	189.7	121.3	1.3	0.3	9.1	2.9	30.6	1.1	6.6	261.6	45.4	274.0	292.5	319.1
Estimated	100	0.1	1.3	2.5	10.3	134.4	52.7	0.7	0.0	3.9	0.8	14.4	0.5	1.6	179.9	13.9	182.4	177.2	192.7
	250	0.1	1.5	4.4	12.5	161.4	86.1	1.0	0.1	6.2	1.7	20.9	0.7	3.8	214.3	26.1	220.4	232.9	254.0
	500	0.2	1.6	5.6	13.9	179.0	107.9	1.2	0.1	7.9	2.4	25.7	0.9	5.3	238.7	36.0	247.9	269.6	294.2
	750	0.2	1.7	6.2	14.7	188.3	119.5	1.3	0.2	8.7	2.7	28.4	1.0	6.1	252.3	42.0	263.6	289.2	315.6
	1000	0.2	1.7	6.6	15.3	194.5	127.2	1.4	0.3	9.4	2.9	30.3	1.1	6.6	261.6	46.4	274.5	302.2	329.8

Simulated: calculated by solving for the appropriate quantiles in simulated data for each province.

Estimated: calculated by plugging in the parameters estimates from Table 3.5 in eq. (3.3.5).

Table 3.7: $\text{PML}_{1/x}$ values (in $ billions) for the insurance claims, based on 100,000 simulated years.

	x	NL	PE	NS	NB	QC	ON	MB	SK	BC	YT	NT	AB	NU	East	West	OSFI	Pearson	Kendall
Simulated	100	0.0	0.0	0.1	0.9	12.7	5.6	0.1	0.0	0.1	0.0	0.0	0.0	0.0	17.4	0.3	17.4	17.2	18.6
	250	0.0	0.1	0.2	1.6	22.3	8.2	0.2	0.0	0.3	0.0	0.3	0.0	0.1	28.7	1.1	28.8	28.8	31.1
	500	0.0	0.1	0.3	2.1	28.2	10.3	0.2	0.0	0.7	0.1	1.2	0.1	0.2	36.3	2.0	36.6	36.4	39.4
	750	0.0	0.1	0.4	2.3	30.0	11.7	0.2	0.0	0.9	0.1	1.8	0.1	0.3	39.4	2.9	39.9	39.3	42.6
	1000	0.0	0.1	0.4	2.5	31.1	12.7	0.2	0.0	1.2	0.2	2.6	0.1	0.4	40.9	3.4	41.6	41.3	44.9
Estimated	100	0.0	0.1	0.1	0.9	12.4	5.8	0.0	0.0	0.1	0.0	0.0	0.0	0.0	17.5	0.3	17.5	17.0	18.4
	250	0.0	0.1	0.2	1.6	21.7	8.2	0.1	0.0	0.3	0.0	0.2	0.0	0.1	27.9	0.9	28.0	28.2	30.5
	500	0.0	0.1	0.3	2.1	27.6	10.5	0.2	0.0	0.6	0.1	1.2	0.1	0.2	35.1	1.8	35.4	35.9	38.9
	750	0.0	0.1	0.4	2.4	30.6	12.0	0.2	0.0	0.8	0.1	1.9	0.1	0.3	39.1	2.7	39.6	40.2	43.6
	1000	0.0	0.1	0.4	2.6	32.5	13.2	0.3	0.0	1.0	0.2	2.5	0.1	0.4	41.9	3.6	42.6	43.2	46.9

Simulated: calculated by solving for the appropriate quantiles in simulated data for each province.

Estimated: calculated by plugging in the parameters estimates from Table 3.5 in eq. (3.3.5).

3.5 Discussion

Canada has elevated seismic risk due to the presence of urban population in zones with high seismic activities, such as in Eastern and Western Canada. Estimation of the financial damage is relevant to insurance companies, governmental institutions and homeowners. Assessment of the spatio-temporal models by means of goodness-of-fit tests can be performed through residual analysis methods. In this paper, we fit two different models to the spatio-temporal point pattern for significant Canadian earthquakes. We extend the use of residual analysis on Voronoi polygons by calculating deviance Voronoi residuals, which provide a very useful tool to compare fitted models and identify locations where one model is superior to the other. We also create an earthquake financial losses estimation tool for Canada by relying on building information, their replacement costs and earthquake damage probability matrices. Additionally, insurance policy terms and market information are used to estimate insurance claim values. A more interpretable approach is suggested to calculate the county-wide PML by relying on the correlation between neighboring provinces. A large simulation of 100,000 years of earthquakes is performed, where we obtain parameter estimates for the tail behavior by using extreme value theory techniques. The results displayed in this article can be improved further by obtaining data on the non-residential buildings in Canada and more detailed information on earthquake insurance penetration and policy terms in Canadian municipalities. Based on our methodology and parameters, a significant earthquake that occurs at a rate of 1-in-500 years in Québec can cause financial damages of around $180 billion. Yet, earthquake insurance claims that occur at a rate of 1-in-500 years are only around $28 billion. Some of the $152 billion in uncovered losses can be partially covered by homeowners, but the government may have to intervene, especially to repair damages in infrastructure. This represents considerable government expenditures, for example, it is nine-fold the budgeted COVID-19 support and recovery measures for the years 2020-2024. Covering $152 billion

would double the planned total expenditures for 2021-2022 (Gouvernement du Québec 2021) and lead to a major deficit. Analogous comparisons show that governments in Ontario and other provinces are also ill-prepared for earthquake relief and that there is a need for further insurance market penetration, especially in Eastern Canada.

Supplementary Material

An open-source interactive web application is provided in https://robabairakdar.shinyapps.io/shiny/. The application allows the user to simulate multiple significant earthquakes, with different random moment magnitudes, in a chosen geographical location in Canada. The insurance market penetration and insurance terms are chosen by the user and their resulting insurance claims are calculated. This web application can provide insurers with a simulated value of the expected financial losses in case of an occurrence of a significant earthquake in areas where they have exposure or plan to sell new earthquake policies. It also provides the simulated isoseismal map, which contains information on the exposure and losses for the affected CSDs.

Chapter 4

Random Forests for Wildfire Insurance Applications

4.1 Introduction

Wildfires burn around 423 million hectares of the global land surface annually, which represents approximately 3% of the global vegetation land (Giglio et al. 2018). Under very dry conditions, such as extreme heat combined with very low precipitations, a natural spark or human activity can ignite a wildfire. The effect of climate on wildfire risk varies depending on the location and time due to the spatial and temporal climate diversity, spatial variation of vegetation, fire management and prevention policies, and the interaction between these variables. The 2016 wildfires in Alberta (Fort McMurray) and the 2017-2021 wildfires in California generated billions of dollars in insured losses and alerted the public and insurers to the threats caused by wildfires. Knowledge of the spatial likelihood of burning and the potential financial impact of wildfires can provide a great tool for wildfire risk management.

Although fire activity in North America fluctuates annually, recent years witnessed an

increase in the annual number and size of forest fires in some locations compared to the past few decades. In fact, the 2017 and 2018 wildfires in British Columbia are the largest wildfires to be recorded in the area since the 1950s, with a total of 1.2 and 1.4 million hectares burned, respectively (Hanes et al. 2019). The Dixie fire in 2021 is the second-largest wildfire in the history of California, with around 0.4 million hectares burned (CALFIRE 2022).

Wildfire losses fluctuate considerably, depending on the location of events. Between 2000-2019, 26% of the world's wildfires and 69% of the economic losses due to wildfires occurred in the United States of America (USA), with the most damage in California (Yaghmaei 2020). In California, numerous insurers have stopped writing homeowners' insurance policies in fire-prone areas, resulting in a transfer of the insurance business to expensive specialty markets (Groom 2015). The Canadian Interagency Forest Fire Center (CIFFC) reports 6,317 wildfires across Canada in 2021, with more than 4.2 million hectares of burned land (CIFFC 2021). The International Disaster Database reports that North America suffered from $68.8 billion in total damages (adjusted to inflation) due to wildfires over the period between 2000-2019, with only 64% of those losses being insured (EM-DAT 2020).

Spatial wildfire risk assessment and fire spread models are growing fields with applications in fire-related decision making. Simulation-based fire spread algorithms were introduced to mitigate the damage caused by wildfires; see, for example, Van Wagtendonk (1996). Spatial wildfire models in the literature usually focus on a limited geographical study area; see, for example Finney et al. (2011) (continental USA), Atkinson et al. (2010) (Tasmania, Australia), Chuvieco et al. (2010) (Spain), Massada et al. (2009) (Northwestern Wisconsin, USA), and Van Wagtendonk (1996) (Sierra Nevada, USA), however, they are not created for actuarial purposes. Adequate estimation and prediction of future wildfire insured losses is challenging since it requires information on buildings. Moreover, constant developments and improved construction in wildfire-prone areas may not be consistent with historical

losses. As such, insurers are typically buying private complex wildfire insurance risk models, which may provide results that are difficult to interpret and replicate. Spatial modeling in the actuarial literature usually focuses on claims data; see, for example Shi and Shi (2017) for spatial frequency and severity modeling of auto-insurance claims, however, one may not always have sufficient claims data that covers the geographical region of interest. In the context of natural catastrophic events, such as wildfires, it is more suitable to follow a natural catastrophe risk modeling decomposition approach, i.e. identify the hazard, vulnerability and exposure; see Mitchell-Wallace et al. (2017) for details. There is wildfire risk at the intersection of those three components, such that a vulnerable property is subject to wildfire risk due to its location and characteristics.

This article provides a quantitative approach that considers the spatial variation in climate, land and demographic factors that drive wildfire occurrence in North America. We propose a wildfire risk model specifically tailored for USA and Canada, separately, splitting for each cause of ignition: man-made and natural. Our goal is to provide an open-source model, for pricing and underwriting P&C insurance policies. The model takes into account multiple types of wildfire risk features: continuous variables, categorical variables, and geographical coordinates. Several modeling techniques are used, such as GLMs and statistical learning algorithms that rely on decision trees. Combining numerous decision trees in an ensemble, such as a random forest, provides high predictive performance with an ability to discover interaction effects between the predictors. Tree-based machine learning models offer strong predictive capabilities, in addition to their interpretability and ease of explanation of the importance of the predictors, unlike neural network. They can also implicitly handle variable correlation, capture variable interactions and non-linear relationships between the predictors. Given the lack of publicly available wildfire claims data for each location in North America, we instead focus on modeling the annual burn probabilities by relying on publicly-available historical wildfire data. The goal of the article is to provide an interpretable model that predicts the annual burn probabilities for any location in North

America and to compare how random forest models perform compared to more commonly used actuarial modeling techniques, such as GLMs. The best fit model is used in a downscaling exercise where we predict the annual burn probabilities at a high resolution, which is applicable for insurance purposes, such as pricing home insurance policies with coverage against wildfires.

The remainder of the article is organized as follows. Section 4.2 explains the data used in the model, such as the historical wildfires data, vegetation information and land use, meteorological and population data. It also classifies the population of North America depending on the wildfire risk. Section 4.3 provides details on the methods proposed to fit the data. It also summarizes the results of the fitted models and compares their predictive capabilities. Section 4.4 provides a detailed analysis of the effect of each input variable. An insurance application is illustrated in Section 4.5 to underwrite and price insurance policies. Finally, Section 4.6 concludes the article.

4.2 Data

In this section, we discuss the data used to build our model. We explain the historical wildfires data and the wildfire risk factors such as climate, land cover, population census, and lightning frequency.

4.2.1 Wildfire occurrence and intensity

The area burned and the frequency of wildfires are relevant metrics to identify trends in wildfire risk and they are also of great importance to estimate insurance losses. Summaries of the historical wildfires in North America are provided in Figure 4.1, where the data is obtained from CFS (2019) and Short (2017). Data for Canadian wildfires range from 1975 until 2019. 2016 witnessed 1.3 million hectares burned - the Fort McMurray wildfire being

responsible for 0.6 million. The last five years of data (2015-2019) had 29,000 wildfires causing a total of 13 million hectares burned, with an average fire size of 450 hectares. Data for the USA wildfires range from 1992 until 2015. The last five years of data (2011-2015) had 370,000 wildfires resulting in a total of 15 million hectares burned, with an average fire size of 40 hectares. The increase in the area burned can be attributed to the lengthening of fire seasons due to the increase in spring and summer temperatures, as explained by Westerling et al. (2006). Historical wildfires in the USA compared to Canada contain a substantial number of wildfires that are small in size, however, Canadian wildfires data contain relatively larger fires. This may be explained by the differences in reporting of wildfires by the fire agencies.

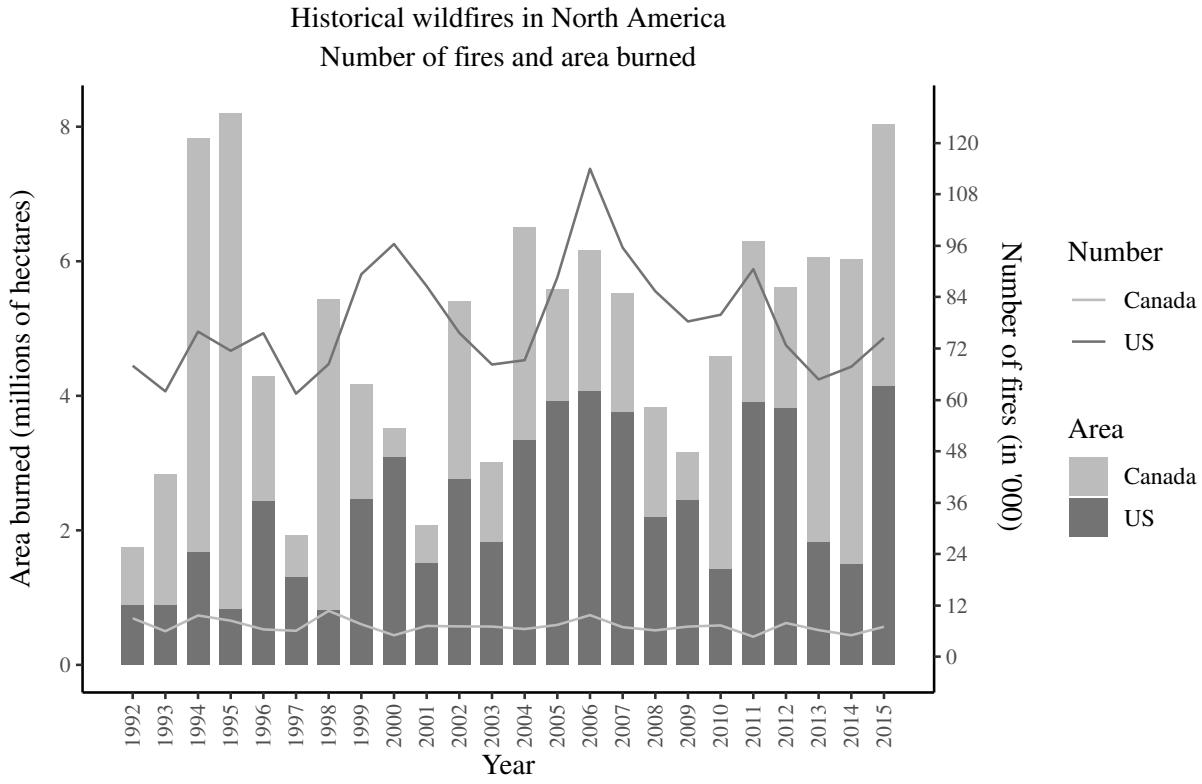


Figure 4.1: The annual number of wildfires (lines) and area burned (bars) in North America. Each bar shows the split of the total area burned over the USA and Canada.

Lightning strikes are the leading cause of large wildfires, which is typically due to prolonged dry conditions, high temperatures, and excessive dry fuel. However, in densely populated

areas, man-made sparks can be responsible for a large number of ignitions due to human-related activities, such as arson, campfires, equipment use, fireworks, smoking, etc. Figure 4.2 compares the annual number and size of wildfires of all sizes caused by lightning strikes and human ignitions, where we observe that natural wildfires are significantly less frequent compared to man-made wildfires, but they tend to burn more land and cause more damage.

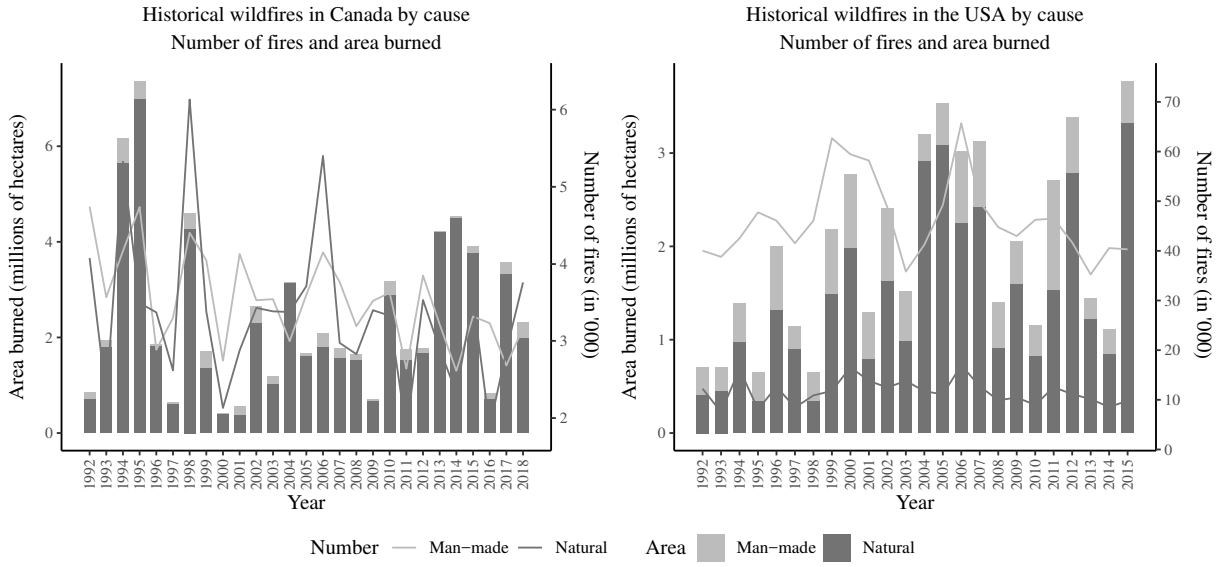


Figure 4.2: Annual number of wildfires (lines) and area burned (bars) by cause of wildfire in North America. Each bar shows the split of the total area burned by the cause of the wildfire; i.e. man-made vs. natural wildfires.

Numerous small wildfires may not necessarily result in a significant increase in the area burned. Accordingly, it may be more appropriate to only consider large wildfires as they are more likely to cause material damage. The Canadian large wildfire database includes information on all wildfires larger than 200 hectares (Stocks et al. 2002). The large wildfires in Canada represent 4% of the count of all wildfires in Canada and 98.8% of the total area burned in the observed period. 84% of the Canadian large wildfires are ignited by lightning strikes and they contribute to 91% of the total area burned by all large wildfires in Canada. By applying the same threshold of 200 hectares, less than 1% of the count of wildfires in the USA is considered large and they correspond to 88.8% of the total area burned. 44.6% of the large wildfires in the USA are ignited by natural reasons and they contribute to

67.8% of the total area burned.

Figure 4.3 maps the historical annual burn probabilities for each cause per grid cell. The chosen resolution is $0.1° \times 0.1°$ where $1°$ is approximately 111 km at the Equator, which is the resolution of the meteorological data. The count of wildfire origins that occurred in each cell is calculated for the entire duration of each dataset, 45 years for Canada and 24 years for the USA. Each year is coded as binary because only a few grid cells have more than one wildfire in a given year. The annual burn probability is obtained by dividing by the number of years of data. Due to the different natures of the maps, we choose to split our North American wildfire model by country and by cause. Wildfire risk is heterogeneous in space, which means that observations are not identically distributed, i.e. in each grid cell, there might be a different distribution. In addition, one could have spatial dependence in between the random variables, i.e. if there is a claim in a grid cell, then it is highly likely there is a claim in the neighbouring grid cell.

4.2.2 Predictors/Determinants of wildfires

Table 4.1 summarizes the wildfire risk predictors explained in this section and used throughout this paper. All meteorological data is obtained from the Copernicus Climate Change Service Climate Data Store (CDS) ERA5 reanalysis data at a spatial resolution of $0.1° \times 0.1°$ (CDS 2019). Monthly-mean averages for the observed period are obtained for the following variables: (1) the air temperature at 2m above the surface of the Earth, (2) the horizontal speed of air moving towards the east and the north at a height of 10m above the surface of the Earth, and (3) the total precipitation from the accumulated liquid and frozen water, including rain and snow that falls to the surface of the Earth. The data is then averaged by fire season, where fire season is defined to be March to September for the USA wildfires, May to September for Canada's natural wildfires, and April to October for Canada's man-made wildfires. The fire seasons are obtained from the historical wildfire

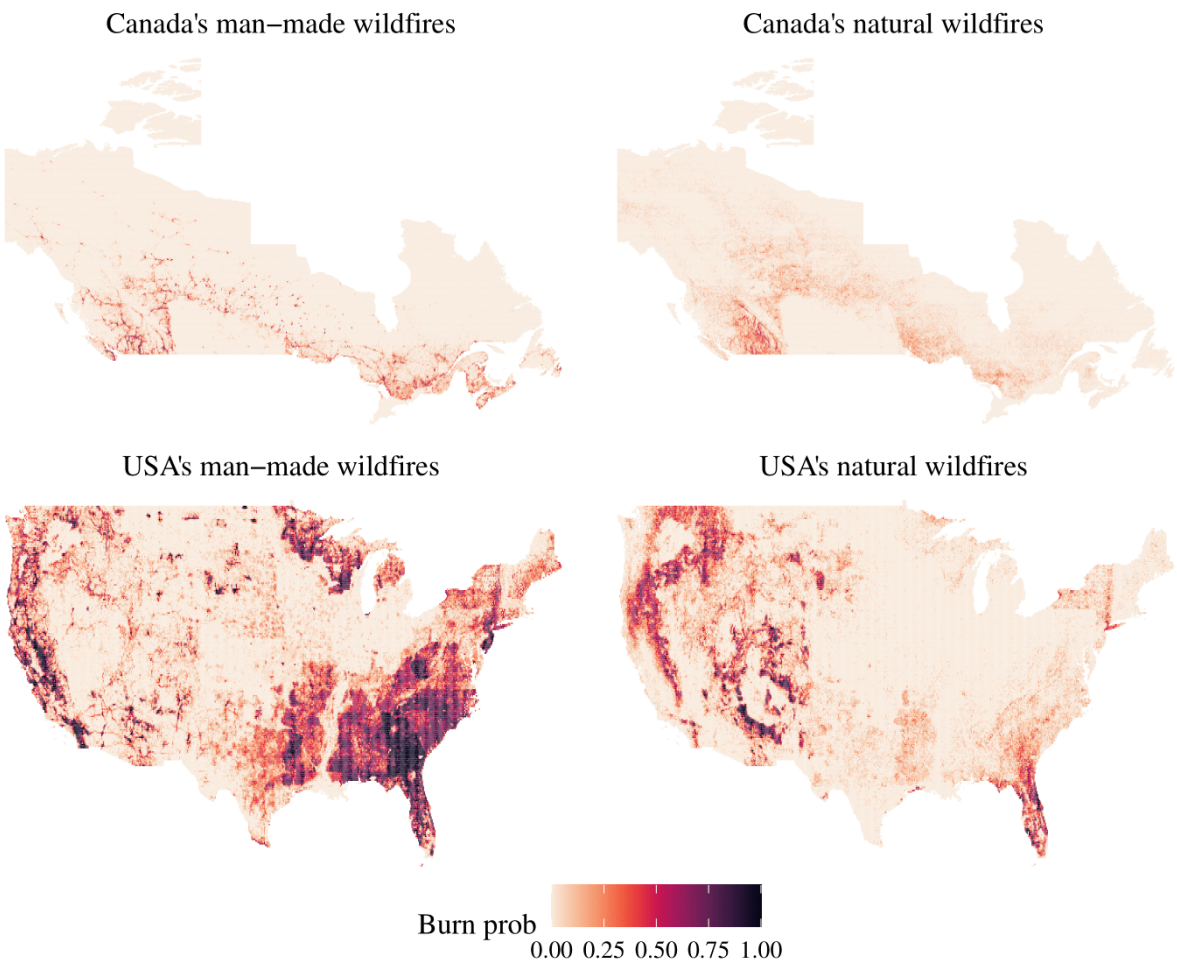


Figure 4.3: Observed annual burn probabilities per grid cell. The maps are split by country and cause of wildfire.

data by observing the monthly seasonality for each wildfire cause. All other covariate information are aggregated/re-scaled to $0.1° \times 0.1°$ to unify the resolution and are stored as raster images, which are pixel-based files that contain unique information per cell.

We use the 2015 North American Land Cover 30m dataset (NALCMS 2020), which explains the material features of the Earth's surface at a 30m spatial resolution based on Landsat-7 imagery. To unify the resolution with the other covariates, the percentage of each land type is calculated for each $0.1° \times 0.1°$ grid cell, hence, converting the 19 types of land cover from a categorical variable to 19 continuous variables that range from 0 to 1. The world digital elevation model, created from a digital database of land and sea-floor elevations on a $1°/12$

resolution, is used as a predictor that represents elevation (NOAA 2016). The global map of lightning frequency contains the number of strikes/km2/year on a $0.5° \times 0.5°$ grid. In North America, Florida has the highest annual lightning strike rate, with an average of 59 strikes/km2/year (Cecil, Buechler, and Blakeslee 2014).

Population count data is obtained from the fourth version of the Gridded Population of the World collection, which provides a smoothed distribution of the human population on a continuous global raster surface at a resolution of $1°/120 \times 1°/120 \approx 1$ km $\times 1$ km (CIESIN 2017). The dataset is created from census data collected between 2005 and 2014 and then extrapolated to estimate the population count/density for the years 2000, 2005, 2010, 2015, and 2020. For this article, we use the raster for the population count in 2005 as a possible predictor and the population count in 2020 is used for predictions.

4.2.3 Population at risk

To analyze the population exposed to wildfire risk, we classify the population in each state and province into low, medium or high risk based on the observed burn probabilities of natural or man-made wildfires. We define low risk to be an annual likelihood $< 10\%$, medium risk is between $10 - 40\%$ and high risk are for grid cells above 40%. Table 4.2 summarizes the results. Overall, 8% of Canadian residents are at high risk of man-made wildfires, compared to 33% of US residents. While no Canadian residents are located in high risk zones for natural wildfires, 5% of the USA population resides in zones that are at high risk for natural wildfires.

4.3 Wildfire Occurrence Models

This section explains the models used in this paper and then provides the results of the fitted models and compares their predictive capabilities. We are interested in modeling

Table 4.1: Definitions of the modeling predictors.

Variable	Definition
Longitude	The longitude at the center of the grid cell
Latitude	The latitude at the center of the grid cell
LandCover_1	Percentage of the grid cell land covered by temperate or sub-polar needleleaf forest
LandCover_2	Percentage of the grid cell land covered by sub-polar taiga needleleaf forest
LandCover_3	Percentage of the grid cell land covered by tropical or sub-tropical broadleaf evergreen forest
LandCover_4	Percentage of the grid cell land covered by tropical or sub-tropical broadleaf deciduous forest
LandCover_5	Percentage of the grid cell land covered by temperate or sub-polar broadleaf deciduous forest
LandCover_6	Percentage of the grid cell land covered by mixed forest
LandCover_7	Percentage of the grid cell land covered by tropical or sub-tropical shrubland
LandCover_8	Percentage of the grid cell land covered by temperate or sub-polar shrubland
LandCover_9	Percentage of the grid cell land covered by tropical or sub-tropical grassland
LandCover_10	Percentage of the grid cell land covered by temperate or sub-polar grassland
LandCover_11	Percentage of the grid cell land covered by sub-polar or polar shrubland-lichen-moss
LandCover_12	Percentage of the grid cell land covered by sub-polar or polar grassland-lichen-moss
LandCover_13	Percentage of the grid cell land covered by sub-polar or polar barren-lichen-moss
LandCover_14	Percentage of the grid cell land covered by wetland
LandCover_15	Percentage of the grid cell land covered by cropland
LandCover_16	Percentage of the grid cell land covered by barren lands
LandCover_17	Percentage of the grid cell land covered by urban
LandCover_18	Percentage of the grid cell land covered by water
LandCover_19	Percentage of the grid cell land covered by snow and ice
Elevation	Height above sea level
Population	Count of individuals who reside in the grid cell
LightningFrequency	The annual frequency of lightning strikes per grid cell
Temperature_FireSeason	The average daily temperature during the fire season
Temperature_NotFireSeason	The average daily temperature in days falling outside the fire season
TotalPrecipitation_FireSeason	The average daily total precipitation during the fire season
TotalPrecipitation_NotFireSeason	The average daily total precipitation in days falling outside the fire season
EastwardWind_FireSeason	The average speed of air moving towards the east during the fire season
NorthwardWind_FireSeason	The average speed of air moving towards the north during the fire season

Table 4.2: Wildfire risk classification by population. The states and provinces are ordered in descending order of their count of wildfires. The top 5 provinces in Canada and the top 10 states in the USA are shown.

Man-made					Natural				
Province	Number of wildfires ('000)	Low risk (%)	Medium risk (%)	High risk (%)	Province	Number of wildfires ('000)	Low risk (%)	Medium risk (%)	High risk (%)
Canada	179	75	17	8	Canada	153	97	3	0
British Columbia	42	49	25	26	British Columbia	51	79	19	2
Ontario	30	93	5	2	Ontario	28	99	1	0
Alberta	28	92	4	4	Alberta	25	98	2	0
Québec	27	56	34	10	Saskatchewan	11	98	2	0
Saskatchewan	13	90	3	7	Québec	11	100	0	0
USA	1105	43	25	33	USA	274	85	10	5
Georgia	149	6	14	80	Oregon	32	84	10	5
California	97	27	33	40	Arizona	29	90	6	4
Texas	93	37	49	14	California	27	93	6	1
North Carolina	80	2	9	89	Florida	22	41	34	25
Mississippi	60	6	17	77	Idaho	21	83	15	1
Alabama	55	0	23	77	Colorado	20	80	15	5
Florida	54	12	17	70	New Mexico	18	70	26	4
South Carolina	49	1	12	87	Montana	18	66	28	6
Minnesota	37	59	17	24	Utah	17	64	30	6
Tennessee	29	40	34	26	Nevada	11	78	19	3

Low risk: annual likelihood $< 10\%$, medium risk: annual likelihood 10 - 40%, and high risk: annual likelihood $> 40\%$.

N, the random number of wildfires for each location in North America over a grid of $0.1^\circ \times 0.1^\circ$, over a time period t; 45 years for Canada and 24 years for the USA. Due to the significant natural climate variability, we use as much available data over time as possible to cover most possible climate scenarios. The annual burn probability per grid cell is N/t. Following the discussion of determinants of wildfires in Section 4.2.2, the predictors of the models are: geographical location, percentage coverage over a grid cell by each land type, elevation, population count, lightning frequency, the mean temperature in fire season, the mean temperature other than fire season, the mean precipitation in fire season, the mean precipitation other than fire season, the mean eastward wind speed in fire season and the mean northward wind speed in fire season. See Table 4.1 for the definitions of the predictors.

Wildfire occurrence models are conditional on meteorological, topographical and socioeconomic variables, as such, any trend in wildfire occurrence is implicitly captured by the trend in the covariates.

4.3.1 Methodology

We compare GLMs (McCullagh 1984), regression decision trees (Breiman et al. 2001) and random forests (Breiman 2001). Even though interpreting GLMs is quite simpler than tree-based models, yet the latter can capture interactions easily and model non-linear relationships between predictors. Decision trees provide predictions that are easily interpreted, however they are known to have high variance, where small changes in the data can generate different trees and hence different predictions for some observations. This limitation is taken care of by using random forests, which relies on numerous decision trees, thus minimizing the variance.

To fit the models and compare their predictive power, we first randomly partition the data into train (90%) and hold-out test (10%) datasets. We apply repeated 10-fold

cross-validation (CV) with 5 repetitions to train the candidate models and tune the hyperparameters of the tree-based models. As a model validation technique, CV allows us to assess how the performance of our models can generalize to other independent datasets. In each iteration of the CV, we use 9 different folds out of the 10 folds to train our model, and we test on the remaining fold. We evaluate the performance of the tree-based models over an extensive grid search of possible values of the hyperparameters. The best set of hyperparameters are the ones that minimize the root mean square error (RMSE); see Bruce and Bruce (2017) for further details. The hold-out test dataset is later used to compare predictions with the actual observed values. The `Caret` (Kuhn 2021) R package is used to tune the hyperparameters, train the models and calculate variable importance.

In the next section, we present the details of the fitted models. First, we model the occurrences of wildfires of any size, split by country and cause of wildfire. Then, we consider large wildfires only and we model them by country.

4.3.2 Models for wildfires of any size

The wildfires are split by ignition cause; man-made or natural. Accordingly, we fit four models to the following data sets:

(i) Canada man-made wildfires,

(ii) Canada natural wildfires,

(iii) USA man-made wildfires, and

(iv) USA natural wildfires.

Figure 4.4 displays the distribution of the response variable in each data set, representing generally balanced datasets. Canada man-made wildfires had a maximum of 44 wildfires in a grid cell over 45 years, i.e. almost a wildfire occurred in approximately the same location

every year. However, Canada's natural wildfires had a maximum of 32 wildfires in a grid cell over 45 years. The USA man-made and natural wildfires each had a maximum of 24 wildfires over 24 years.

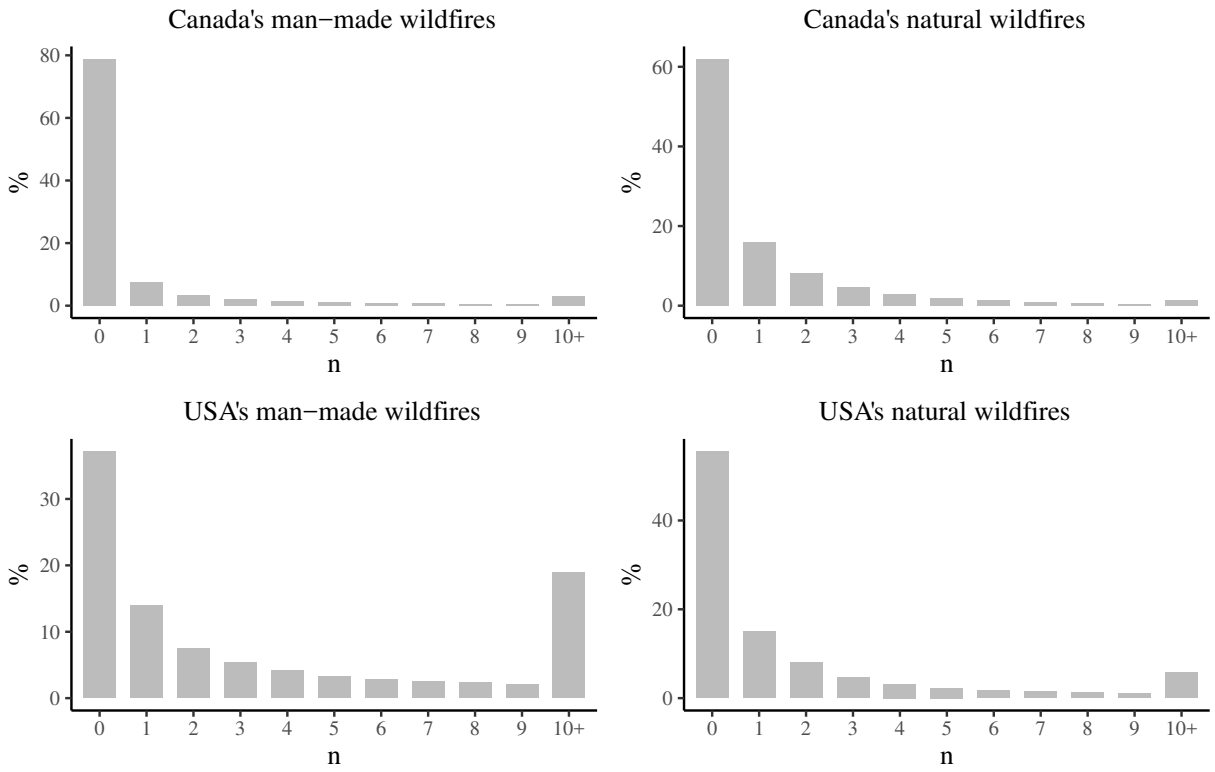


Figure 4.4: The proportion of $0.1° \times 0.1°$ grid cells with n wildfires of any size in 45 years in Canada and 24 years in the USA.

We apply the algorithms discussed in Section 4.3.1 to train the data and tune the hyperparameters of the tree-based models. To assess the predictive accuracy of the models, we calculate the CV root mean squared prediction error and the CV mean absolute prediction error (MAE). To facilitate the comparision between the models for Canada and the USA, standardized versions of the RMSE and MAE, that account for the number of years in each model, are calculated. Thus, both measures are presented in terms of the annual burn probability. For model comparison purposes, we also compare the CV Pearson correlation coefficient between the observed values and the predictions, and we compute the RMSE on the hold-out test dataset. Additionally, we calculate the prediction error

measures for a baseline model, which assumes that the country-wide mean is applied to all grid cells. Table 4.3 summarizes the discussed measures. As expected, the random forest models outperform the other fitted models. We also observe that the random forest models show around half the prediction error of the baseline model.

Table 4.3: Cross-validation prediction errors and Pearson correlation coefficient and hold-out RMSE of the attempted models for fitting wildfires of any size for each country and cause. RMSE and MAE are presented in terms of the annual burn probability.

	Fitted model	Canada		USA	
		Man-made	Natural	Man-made	Natural
CV RMSE	Baseline	0.0706	0.0499	0.2586	0.1525
	GLM - Poisson	0.0617	0.0377	0.2032	0.1178
	GLM - Negative Binomial	0.0706	0.0399	0.2438	0.1264
	Regression Decision Tree	0.0580	0.0435	0.2179	0.1244
	Random Forest	0.0426	0.0267	0.1144	0.0722
CV MAE	Baseline	0.0343	0.0305	0.2036	0.0996
	GLM - Poisson	0.0234	0.0199	0.1296	0.0681
	GLM - Negative Binomial	0.0264	0.0201	0.1488	0.0702
	Regression Decision Tree	0.0244	0.0252	0.1565	0.0755
	Random Forest	0.0156	0.0148	0.0745	0.0407
CV Correlation	GLM - Poisson	0.4985	0.6582	0.6212	0.6372
	GLM - Negative Binomial	0.4555	0.6276	0.5219	0.6089
	Regression Decision Tree	0.5689	0.4873	0.5387	0.5788
	Random Forest	0.7988	0.8445	0.8980	0.8819
Hold-out RMSE	Baseline	0.0703	0.0501	0.2604	0.1525
	GLM - Poisson	0.0597	0.0371	0.2082	0.1191
	GLM - Negative Binomial	0.0721	0.0391	0.2524	0.1279
	Regression Decision Tree	0.0598	0.0457	0.2223	0.1290
	Random Forest	0.0416	0.0261	0.1154	0.0697

The random forest models have the strongest predictive power, and hence they are chosen as the optimal models such that all the analysis thereafter is performed on them. As a comparison tool against historical wildfire records, Figure 4.5 provides a visual representation of the differences between the predictions of the random forest models and the observed events for each grid cell to highlight zones where the model is lacking predictive power. To facilitate the comparison between the four maps, the differences are computed

in terms of the annual burn probability.

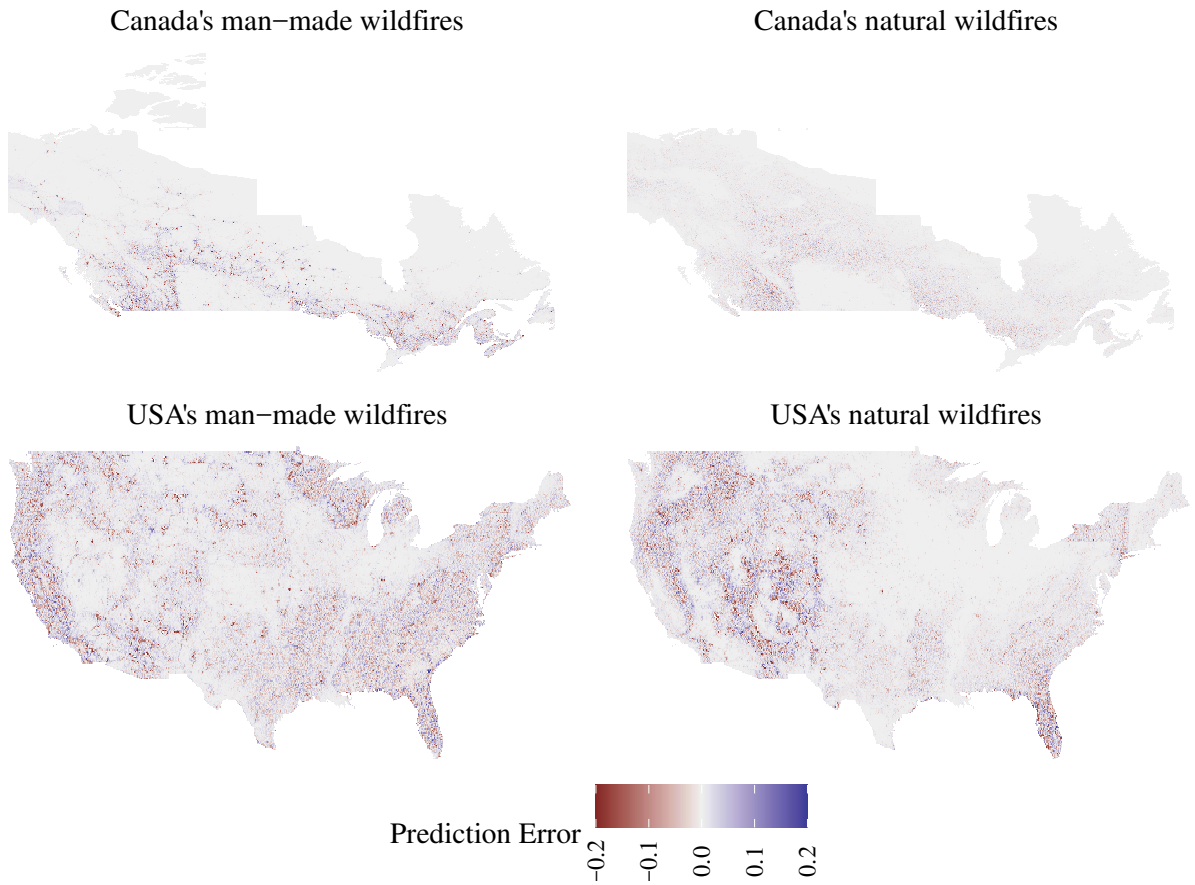


Figure 4.5: Difference between the predicted and actual annual burn probabilities of wildfires of any size.

Figure 4.6 displays the variable importance for each random forest model, which represents the percentage increase in the mean square error of the random forest model when the data for that variable are randomly permuted (Breiman et al. 2001). The analysis shows that for man-made wildfires in Canada and the USA, the most important predictors are population, urban land, water and northward wind speed. For natural wildfires, there are no common important predictors between the countries.

Figure 4.7 compares the predicted annual burn probabilities to the observed burn probabilities in the hold-out test dataset, where we notice a strong linear relationship close to the 45° line, except for slight under prediction of high-risks. This is generally acceptable

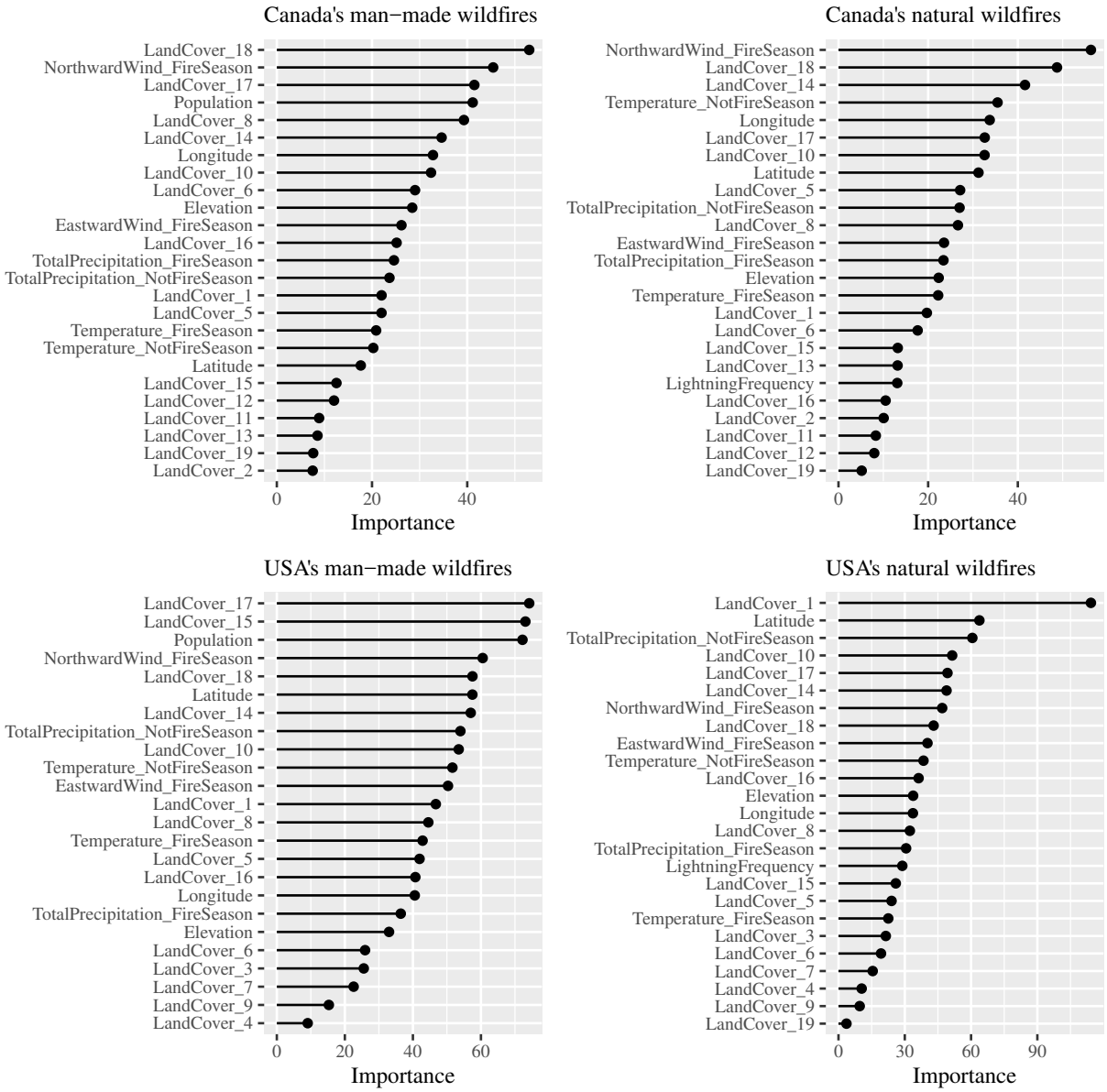


Figure 4.6: Variable importance (percentage mean decrease in accuracy) of the random forest models for wildfires of any size for each country and cause.

due to the difficulty of any model in predicting outliers/extremes. For each observation, prediction intervals are plotted, which represent 95% of the individual predictions from the decision trees composing each random forest model. In fact, around 90% of the observed burn probabilities for the USA wildfires and 95% of the observed burn probabilities for Canadian wildfires fall within the 45^{th} and 55^{th} quantiles of the individual decision tree

predictions.

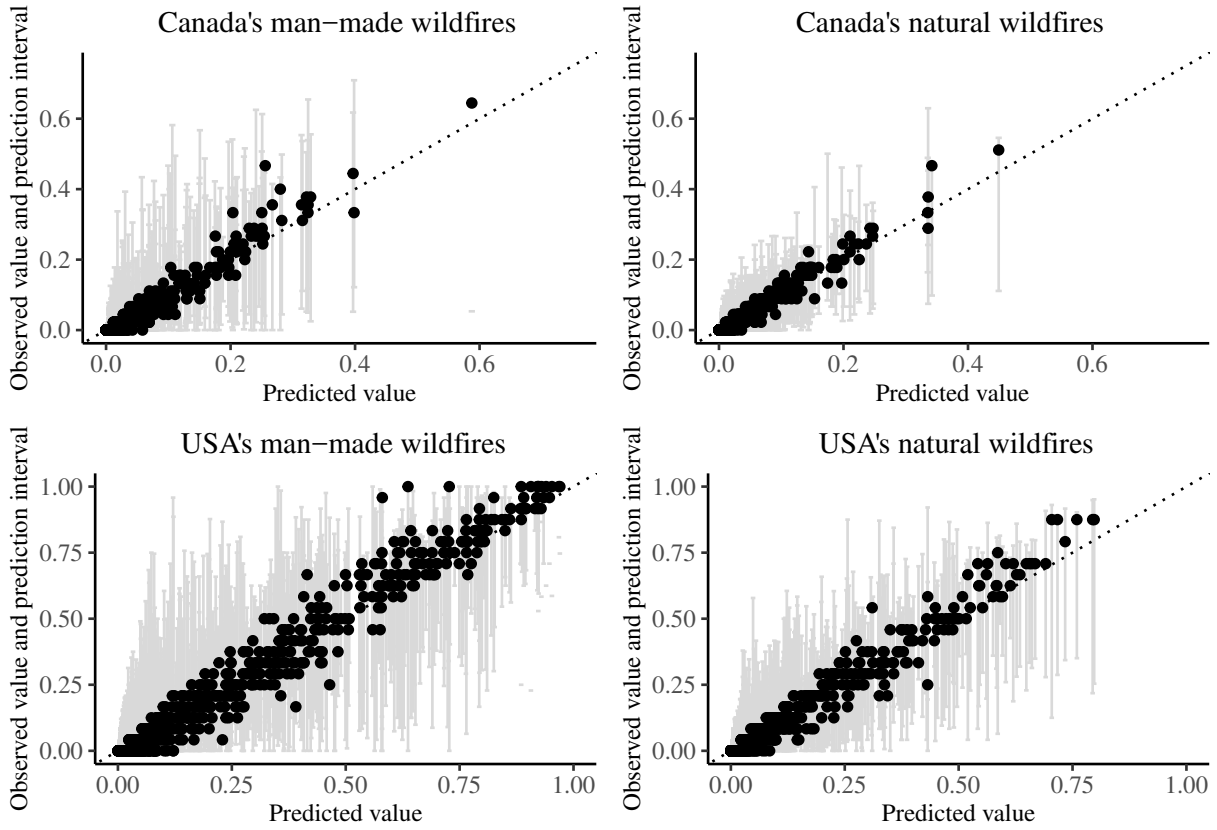


Figure 4.7: Predicted vs observed annual burn probabilities of wildfires of any size, computed on the hold-out test datasets. The vertical bars are the prediction intervals.

4.3.3 Models for large wildfires

Even though large wildfires are small in number, they contribute significantly to the area burned and economic damage. Figure 4.8 displays the distribution of the response variable in each model. There was a maximum of 10 wildfires in a grid cell over 45 years in Canada, compared to 15 large wildfires in the USA over 24 years. We fit two models:

(i) Canada large wildfires, and

(ii) USA large wildfires.

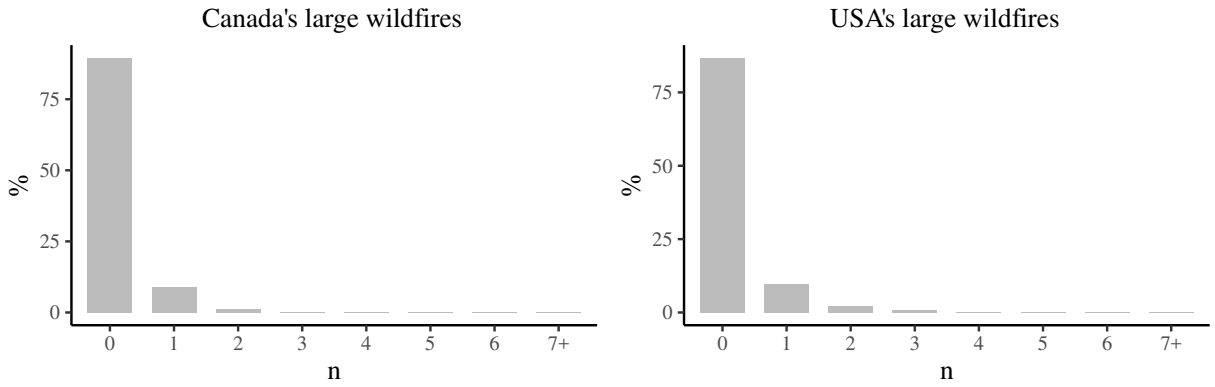


Figure 4.8: The proportion of $0.1° \times 0.1°$ grid cells with n large wildfires in 45 years in Canada and 24 years in the USA.

See Table 4.1 for the predictors of the models. We apply the methods discussed in Section 4.3.1 to train the data and tune the hyperparameters. Table 4.4 provides the CV RMSE, MAE and Pearson correlation coefficient of the fitted models, in addition to the RMSE computed on the hold-out test dataset. Both RMSE and MAE are presented in terms of the annual burn probability. Similar to the models in Section 4.3.2, the random forest models outperform the GLM and decision tree models. It is observed that the Pearson correlation coefficient of both models is low in comparison with the models for wildfires of any size, discussed in Section 4.3.2, which is attributed to the difficulty in predicting the large observation values.

As shown in Figure 4.8, where we have a large number of grids with no observed large wildfires, it is easier for the model to predict the zeros. Accordingly, we consider hurdle models, first introduced by Cragg (1971), which are pure mixtures of zero and non-zero outcomes. A logistic GLM is responsible for the binary component of whether the outcome is zero or positive. If the outcome is positive, the conditional distribution of the non-zeros is modeled by a zero-truncated count distribution. We apply the same methodology by using two-stages random forest model: the first stage is a classification model that categorizes the observations into "no large fire" vs "large fire" occurrence, while the second stage is a regression model to predict the number of occurrences should at least one large wildfire

Table 4.4: Cross-validation prediction errors and Pearson correlation coefficient and hold-out RMSE of the models for fitting large wildfires for each country. RMSE and MAE are presented in terms of the annual burn probability.

	Fitted model	Canada	USA
CV RMSE	Baseline	0.0086	0.0244
	GLM - Poisson	0.0082	0.0232
	GLM - Negative Binomial	0.0082	0.0233
	Regression Decision Tree	0.0084	0.0235
	Random Forest	0.0079	0.0204
CV MAE	Baseline	0.0048	0.0136
	GLM - Poisson	0.0044	0.0122
	GLM - Negative Binomial	0.0044	0.0122
	Regression Decision Tree	0.0046	0.0125
	Random Forest	0.0042	0.0102
CV Correlation	GLM - Poisson	0.2714	0.2972
	GLM - Negative Binomial	0.2694	0.2956
	Regression Decision Tree	0.2142	0.2657
	Random Forest	0.3760	0.5471
Hold-out RMSE	Baseline	0.0082	0.0241
	GLM - Poisson	0.0079	0.0229
	GLM - Negative Binomial	0.0079	0.0229
	Regression Decision Tree	0.0080	0.0234
	Random Forest	0.0076	0.0198

happens. In the first stage, we apply larger weights on the observations with "large fire" so that they have higher chances of selection in the bootstraps samples for the individual decision trees.

We assess the predictive strength of each model component separately: the hurdle component and the count component. For the hurdle component, we calculate the CV area under the receiver operating characteristic curve (AUC) and accuracy, i.e. the ratio of the correctly predicted observations to the total number of observations. While for the count component, we calculate the CV RMSE and Pearson correlation coefficient of the observed count of large wildfires and predicted count of large wildfires. The RMSE is presented in terms of the annual burn probability. Finally, we compute the same measures on the hold-out dataset. Table 4.5 summarizes the discussed measures. For the hurdle

components, all attempted models perform similarly. By observing the AUC, the random forest models are slightly better, however by comparing the accuracy of the models, the GLM hurdle models seem to perform better. For the count components, the random forest models outperform the GLM hurdle models that fail to predict values in the tail of the distribution.

Overall, the two-stages random forest models have the strongest predictive power, and hence they are chosen as the optimal models for the occurrence of large wildfires in Canada and the USA. All the analysis thereafter for large wildfires is performed on the two-stages random forest models. Figure 4.9 illustrates the difference between the predictions of the random forest models for large wildfires and the observed events for each grid cell, where we observe minimal errors, comparable to the small RMSE values shown in Table 4.5. The values are computed in terms of the annual burn probability.

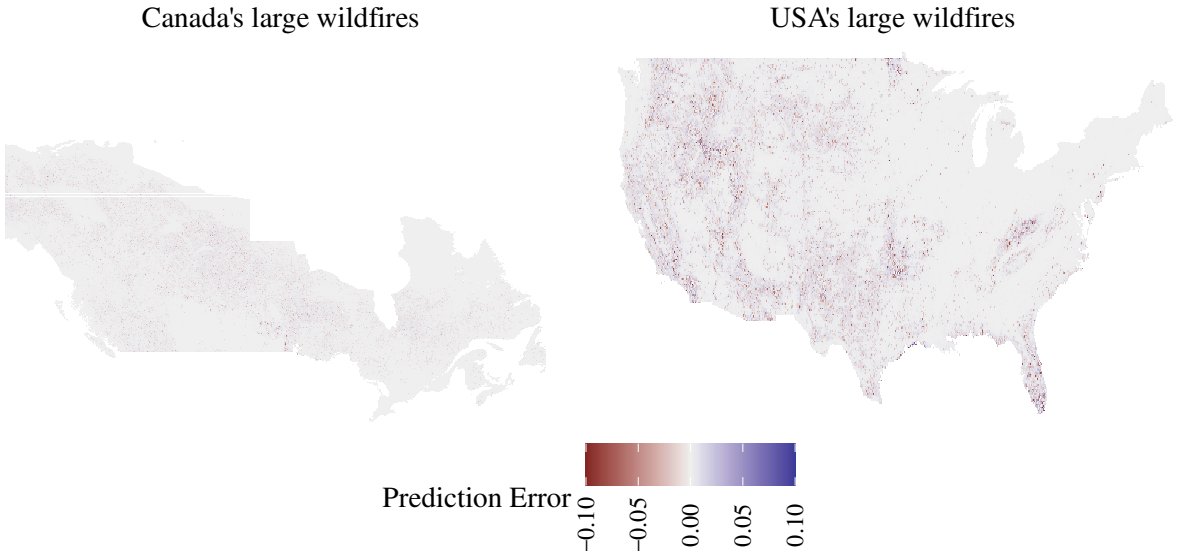


Figure 4.9: Difference between predicted and actual burn probabilities of large wildfires.

Figure 4.10 displays the variable importance for each model in the two-stages random forest models to help identify the most influential variables. The comparison between the variables shows that climate variables are the main drivers behind wildfires becoming large in size.

Table 4.5: Cross-validation and hold-out AUC, Accuracy, RMSE and Pearson correlation coefficient of the attempted hurdle models for fitting large wildfires for each country. RMSE is presented in terms of the annual burn probability.

		Fitted model	Canada	USA
Hurdle Component	CV AUC	GLM - Poisson Hurdle	0.7632	0.7458
		GLM - Negative Binomial Hurdle	0.7631	0.7459
		Random Forest	0.8213	0.8525
	CV Accuracy	GLM - Poisson Hurdle	0.8949	0.8685
		GLM - Negative Binomial Hurdle	0.8949	0.8686
		Random Forest	0.8757	0.8640
	Hold-out AUC	GLM - Poisson Hurdle	0.7494	0.7469
		GLM - Negative Binomial Hurdle	0.7494	0.7469
		Random Forest	0.8170	0.8539
	Hold-out Accuracy	GLM - Poisson Hurdle	0.8959	0.8685
		GLM - Negative Binomial Hurdle	0.8960	0.8686
		Random Forest	0.8749	0.8628
Count Component	CV RMSE	GLM - Poisson Hurdle	0.0233	0.0592
		GLM - Negative Binomial Hurdle	0.0234	0.0593
		Random Forest	0.0098	0.0336
	CV Correlation	GLM - Poisson Hurdle	0.1323	0.2297
		GLM - Negative Binomial Hurdle	0.1317	0.2249
		Random Forest	0.2414	0.4659
	Hold-out RMSE	GLM - Poisson Hurdle	0.0224	0.0581
		GLM - Negative Binomial Hurdle	0.0224	0.0582
		Random Forest	0.0052	0.0197
	Hold-out Correlation	GLM - Poisson Hurdle	0.1264	0.2310
		GLM - Negative Binomial Hurdle	0.1257	0.2264
		Random Forest	0.8809	0.8977

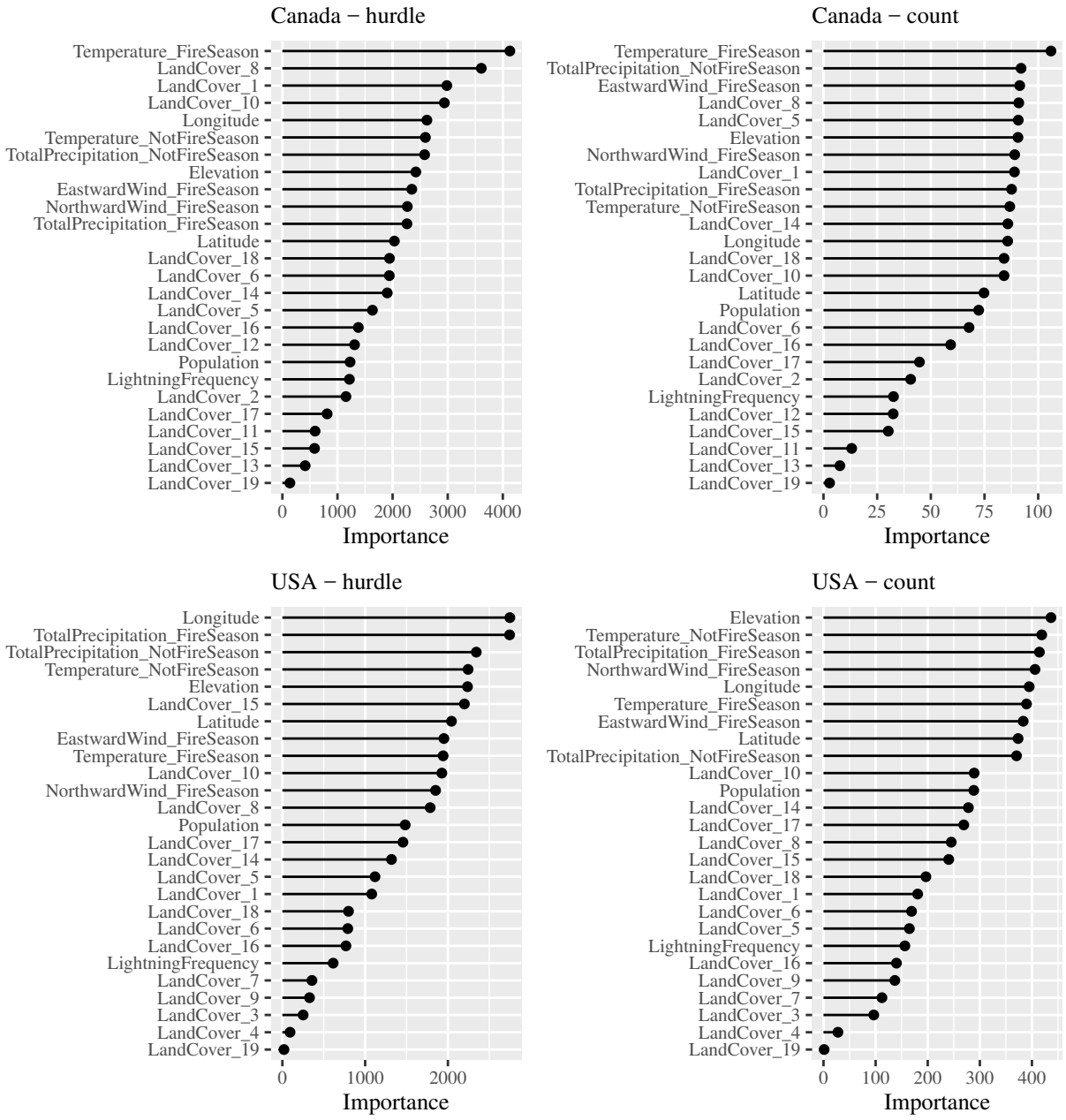


Figure 4.10: Variable importance (percentage mean decrease in accuracy) of the hurdle and count components of the random forest models for large wildfires for each country.

4.4 Sensitivity Analysis

The wildfire random forest models have a large number of input variables, with their importance ranked in Figures 4.6 and 4.10. Note that variable importance does not quantify

Table 4.6: Definitions of the sensitivities.

Sensitivity	Definition
Temp Inc	Increase the mean temperature by 1°.
Temp Dec	Decrease the mean temperature by 1°.
Precip Inc	Increase the mean total precipitation by 10%.
Precip Dec	Decrease the mean total precipitation by 10%.
Lightn Inc*	Increase the annual frequency of lightning strikes per km^2 by 1 strike.
Lightn Dec	Decrease the annual frequency of lightning strikes per km^2 by 1 strike.
Pop Inc	Increase population count and percentage of urban land by 10%.
Pop Dec	Decrease population count and percentage of urban land by 10%.

* An increase in the annual frequency of lightning strikes per km^2 by 1 strike represents an average increase of 103%, 40% and 84% in British Columbia, Alberta and California, respectively.

the effect of changes in the predictors on wildfire risk. In this section, we compute that effect by performing sensitivity tests on some of the predictors, while keeping all other model components unchanged. Table 4.6 contains a list of the performed sensitivities. For each sensitivity, a shock is applied to one (or more) predictor(s), and predictions of the random forest models that are trained in Section 4.3 are calculated. The analyses are evaluated over British Columbia, Alberta and California, which are chosen because of their high level of wildfire risk.

Tables 4.7 and 4.8 provide some summary statistics of the predicted annual likelihood of a wildfire of any size and of large wildfires for each sensitivity, to be compared with the base scenario, where all predictors remain unchanged. The calculated statistics are the mean, standard deviation and percentiles of the burn probabilities over all grid cells in each state and province. More percentiles in the tail are provided for the sensitivities of large wildfires to reflect the behavior in the table. The burn probability of wildfires of any size is the sum of two components; the predicted likelihood of a wildfire due to natural causes multiplied by the historical probability of lightning-caused wildfires and the predicted likelihood of a fire due to man-made causes multiplied by the historical probability of man-made wildfires for each grid cell. The predicted annual likelihood of a wildfire of any size for each cause

is obtained from the random forest models presented in Section 4.3.2. The results of the sensitivities are inline with the expectations; an increase in temperature, a decrease in total precipitation, an increase in lightning activity and human population are all factors that will increase the annual burn probabilities in the study regions. A change in the mean temperature has a stronger effect on the burn probabilities in British Columbia and Alberta compared to California. Even though a change in urban population affects the annual burn probabilities of wildfires of any size, it has a negligible effect on large wildfires because they are more commonly ignited from natural sparks.

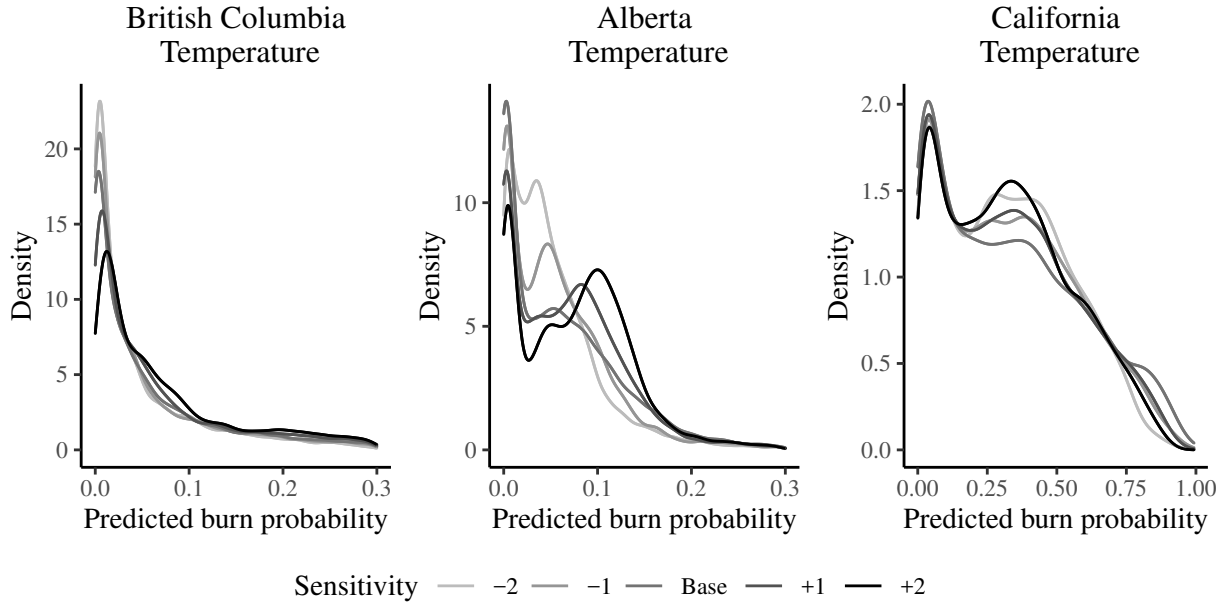


Figure 4.11: Probability density function of the predicted annual burn probabilities of wildfires of any size for the sensitivity tests for the temperature predictors in British Columbia, Alberta and California.

We perform additional sensitivities to compare the distribution of the predicted annual burn probabilities of wildfires of any size over multiple scenarios of change in the mean temperature: $\pm 1°$ and $\pm 2°$. Figure 4.11 compares the densities of the sensitivities performed on the temperature predictor in the three chosen high-risk states and provinces. This figure complements the results from Table 4.7 by providing the full distribution across grid cells. One can see that across all study areas, an increase in temperature reduces the

Table 4.7: Summary statistics of the predicted annual burn probabilities of wildfires of any size for the sensitivity tests.

State / Province	Statistic	Base	Temp Inc	Temp Dec	Precip Inc	Precip Dec	Lightn Inc	Lightn Dec	Pop Inc	Pop Dec
British Columbia	5 percentile	0.0004	0.0014	0.0007	0.0004	0.0006	0.0025	0.0004	0.0004	0.0004
	25 percentile	0.0037	0.0101	0.0050	0.0036	0.0055	0.0126	0.0036	0.0039	0.0037
	Mean	0.0696	0.0783	0.0612	0.0691	0.0709	0.0757	0.0651	0.0703	0.0687
	75 percentile	0.0907	0.1033	0.0795	0.0890	0.0926	0.0979	0.0866	0.0915	0.0900
	95 percentile	0.2840	0.2929	0.2490	0.2791	0.2847	0.2849	0.2473	0.2843	0.2805
	Standard deviation	0.0981	0.0998	0.0864	0.0952	0.0974	0.0956	0.0892	0.0983	0.0962
Alberta	5 percentile	0.0000	0.0003	0.0000	0.0000	0.0003	0.0000	0.0000	0.0000	0.0000
	25 percentile	0.0049	0.0161	0.0042	0.0091	0.0120	0.0051	0.0041	0.0053	0.0048
	Mean	0.0653	0.0709	0.0595	0.0654	0.0699	0.0679	0.0641	0.0659	0.0644
	75 percentile	0.0981	0.1038	0.0854	0.0989	0.1055	0.0998	0.0914	0.0991	0.0974
	95 percentile	0.1955	0.1978	0.1650	0.2008	0.1998	0.1976	0.1874	0.1960	0.1913
	Standard deviation	0.0726	0.0655	0.0620	0.0676	0.0676	0.0723	0.0698	0.0724	0.0705
California	5 percentile	0.0070	0.0107	0.0065	0.0068	0.0080	0.0081	0.0067	0.0072	0.0069
	25 percentile	0.0949	0.1135	0.0908	0.0940	0.1000	0.0990	0.0933	0.1009	0.0948
	Mean	0.3312	0.3364	0.3270	0.3358	0.3389	0.3334	0.3250	0.3333	0.3280
	75 percentile	0.5206	0.5345	0.4991	0.5157	0.5274	0.5298	0.5013	0.5248	0.5125
	95 percentile	0.8094	0.8113	0.7534	0.8003	0.8139	0.8099	0.8088	0.8091	0.8032
	Standard deviation	0.2561	0.2373	0.2365	0.2482	0.2465	0.2551	0.2530	0.2544	0.2531

Table 4.8: Summary statistics of the predicted annual burn probabilities of large wildfires for the sensitivity tests.

State / Province	Statistic	Base	Temp Inc	Temp Dec	Precip Inc	Precip Dec	Lightn Inc	Lightn Dec	Pop Inc	Pop Dec
British Columbia	Mean	0.0022	0.0022	0.0020	0.0022	0.0022	0.0023	0.0021	0.0022	0.0022
	90 percentile	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	95 percentile	0.0236	0.0244	0.0230	0.0233	0.0240	0.0243	0.0233	0.0241	0.0229
	97.5 percentile	0.0248	0.0255	0.0241	0.0240	0.0257	0.0253	0.0243	0.0254	0.0242
	99 percentile	0.0297	0.0311	0.0288	0.0294	0.0305	0.0308	0.0292	0.0310	0.0290
	Standard deviation	0.0072	0.0073	0.0069	0.0072	0.0072	0.0073	0.0071	0.0072	0.0072
Alberta	Mean	0.0027	0.0027	0.0027	0.0026	0.0027	0.0027	0.0027	0.0027	0.0027
	75 percentile	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	90 percentile	0.0230	0.0234	0.0227	0.0226	0.0234	0.0233	0.0228	0.0231	0.0229
	95 percentile	0.0244	0.0253	0.0240	0.0239	0.0253	0.0247	0.0241	0.0246	0.0243
	97.5 percentile	0.0257	0.0264	0.0251	0.0253	0.0265	0.0260	0.0252	0.0258	0.0254
	99 percentile	0.0338	0.0357	0.0315	0.0313	0.0345	0.0342	0.0313	0.0343	0.0331
	Standard deviation	0.0081	0.0081	0.0081	0.0082	0.0081	0.0081	0.0080	0.0081	0.0081
California	Mean	0.0172	0.0181	0.0169	0.0170	0.0173	0.0174	0.0171	0.0173	0.0170
	75 percentile	0.0471	0.0495	0.0465	0.0468	0.0485	0.0482	0.0467	0.0473	0.0468
	90 percentile	0.0603	0.0636	0.0586	0.0591	0.0620	0.0619	0.0586	0.0609	0.0591
	95 percentile	0.0734	0.0742	0.0728	0.0723	0.0739	0.0743	0.0727	0.0741	0.0723
	97.5 percentile	0.0865	0.0890	0.0844	0.0858	0.0883	0.0877	0.0842	0.0887	0.0846
	99 percentile	0.1067	0.1120	0.1048	0.1034	0.1106	0.1104	0.1029	0.1118	0.1042
	Standard deviation	0.0300	0.0301	0.0300	0.0301	0.0300	0.0300	0.0301	0.0300	0.0299

number of grid cells that have small predicted annual burn probabilities and increases the chances of medium and high wildfire risk, and vice versa. Additionally, higher volatility can be observed in Alberta and California, compared to British Columbia.

4.5 Insurance Application

In this section, we aim to illustrate how we can use the models to price and underwrite wildfire risk in an insurance portfolio. The goal is to estimate the expected losses due to wildfire risk, based on the geographical location of the homeowner. A typical catastrophe model is composed of hazard modeling, exposure collection, vulnerability assessment and expected losses calculation; see Mitchell-Wallace et al. (2017) for details. In this article, the hazard element is represented by the rate of wildfire occurrence, which varies across our chosen geographic region. This is reflected in the models built in Section 4.3. Collection of exposure requires the valuation of properties and/or infrastructures at risk. Vulnerability assessment helps in quantifying the relationship between the hazard and the damage by means of a metric, such as damage ratios. Finally, the annual pure premium is calculated in terms of the expected losses in each geographical location.

4.5.1 Hazard

A grid cell of size $0.1^\circ \times 0.1^\circ$ may be too large for some insurance applications, such as pricing exercises, due to the need to understand the risk at a household level. As such, we perform a downscaling exercise where we predict the burn probabilities of a wildfire in a year at a much higher resolution. The chosen resolution is that of the Gridded Population of the World dataset, i.e. $1^\circ/120 \times 1^\circ/120 \approx 1$ km $\times 1$ km grid cells, which is acceptable for insurers' purposes to accurately reflect the location of the insureds. A new high-resolution dataset is created by changing the resolution of all predictors, defined in Table 4.1. For

climate, elevation and lightning strikes predictors, all small grid cells of size $1°/120 \times 1°/120$ within a large grid cell of size $0.1° \times 0.1°$ are assigned the same value. While for land cover predictors, the percentage of each land type is computed for each $1°/120 \times 1°/120$ grid cell. The random forest models in Sections 4.3.2 and 4.3.3 are used to predict the annual burn probabilities of wildfires of any size and of large wildfires per grid cell, by using the new high-resolution dataset.

The left panel in Figure 4.12 shows the burn probabilities of wildfires for each location in British Columbia by using the original dataset of resolution $0.1° \times 0.1°$ and by using the high-resolution dataset of size $1°/120 \times 1°/120$. Smaller grid cells indicate less number of predicted wildfires, and hence smaller burn probabilities. The annual burn probabilities in the small grid cells are scaled such that the overall likelihood of a wildfire is equivalent to that of the lower resolution model, while preserving the weights of the risk of the small grid cells. The right panel in Figure 4.12 shows a zoomed-in map on Lytton Creek and its surrounding region. As observed from the predictions of the original dataset and the downscaling exercise, the region has elevated burn probabilities compared to its surroundings and to other areas in British Columbia. This coincides with Lytton Creek's large wildfire that occurred in 2021 in that area, which caused two fatalities and destroyed around 90% of Lytton village. The models were built on wildfires data up to the year 2019 in Canada, thus a 2021 wildfire is considered an out-of-sample prediction, hence confirming the appropriateness of the predictions of the models.

4.5.2 Vulnerability and exposure

The average claim cost of a wildfire is highly volatile, depending on the location of the wildfire, the construction material of the buildings affected, their location, the fire mitigation measures and the availability of firefighters. Table 4.9 summarizes the average claim cost of the most recent large wildfires that occurred in three high-risk regions, British

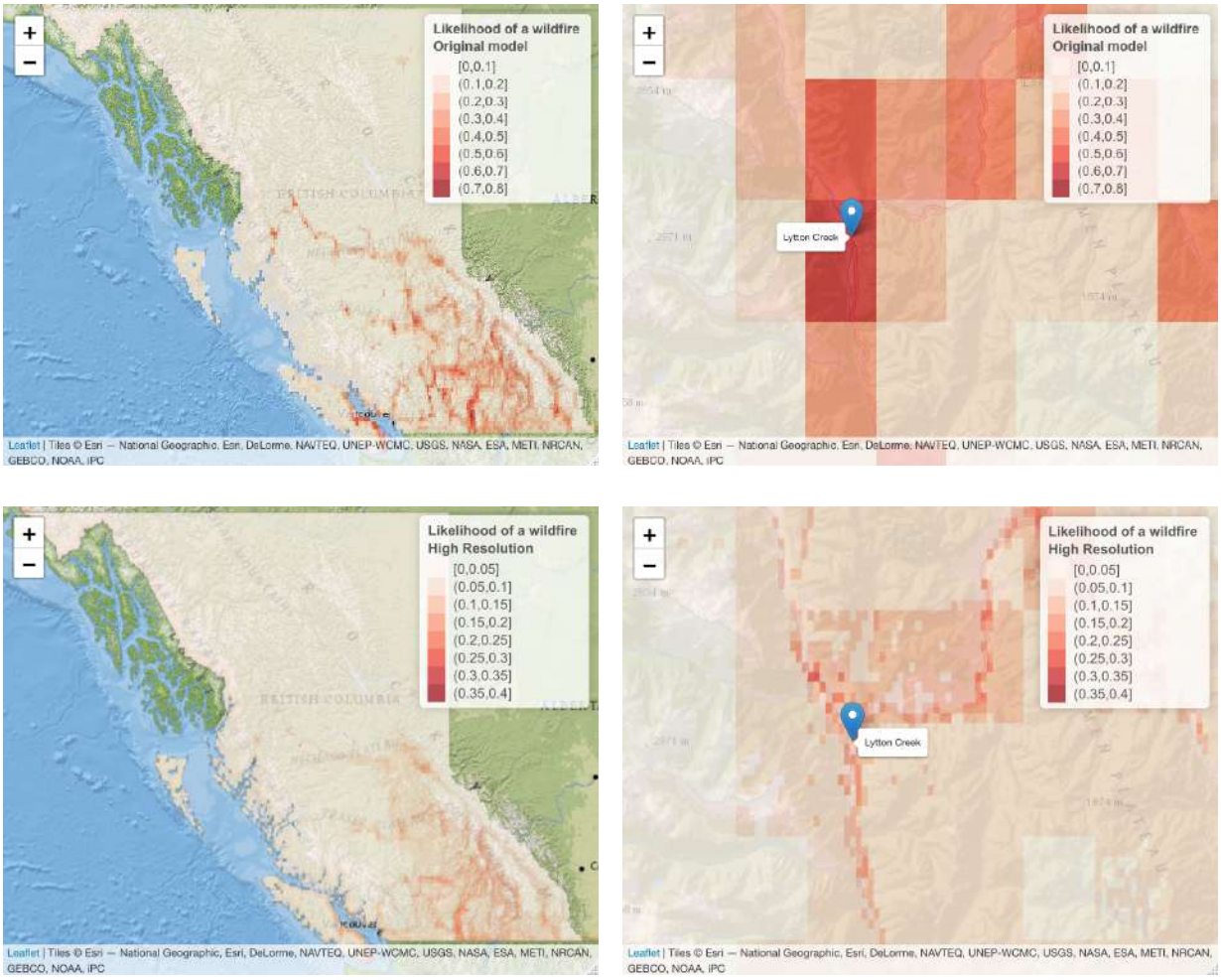


Figure 4.12: Predicted annual burn probabilities of wildfires of any size in British Columbia for the original (top) and high-resolution datasets (bottom), presented for the full province (left) and Lytton Creek (right).

Columbia, Alberta and California. Thus, relying on such limited values to generalize over the whole region can be problematic. The unit of currency throughout is the Canadian dollar for Canadian wildfires and the US dollar for USA wildfires.

Vulnerability assessment for wildfire risk usually requires the help of civil engineers to analyze the characteristics of the properties. Those details are not publicly available, however, such information is typically provided by the customers to their insurers and is used in the calculation of the insured exposure. In this article, we rely on insured exposure data that is provided by CatIQ, Canada's Loss And Exposure Indices Provider

Table 4.9: Historical average claim cost for some memorable large wildfires (CatIQ 2021; IBC 2021; CALFIRE 2021).

State/Province	Wildfire	Year	Average Claim Cost
British Columbia	White Rock Lake	2021	\$55k
	Lytton Creek	2021	\$253k
	Williams Lake	2017	\$15k
	Elephant Hill	2017	\$12k
Alberta	Fort McMurray	2016	\$64k
	Slave Lake	2011	\$110k
California	Multiple wildfires	2020	\$826k
		2019	\$574k
		2018	\$178k
		2017	\$1,116k
		2016	\$116k
		2015	\$958k
		2014	\$30k
		2013	\$60k
		2012	\$113k
		2011	\$52k

(CatIQ 2021). They release annual updates of the Canadian insurance industry exposure database, which is developed from data from the Canadian P&C insurance companies. The 2020 year-end estimates of the personal properties sums insured and number of insured risks against fires is available for each Canada Post Forward Sortation Area (FSA). Depending on the population distribution in the FSA, the exposure value and the number of insured households in an FSA is distributed over each grid cell of size $1^\circ/120 \times 1^\circ/120$. Accordingly, the average exposure value per household is calculated for each grid cell.

4.5.3 Insurance premiums

Pure premium computation requires hazard, vulnerability and exposure calculations to be performed at the homeowner's level. This entails detailed information on the location of the household, to determine its subjection to wildfire risk, in addition to property value and characteristics to evaluate its vulnerability if a wildfire occurs. In this section, we

illustrate how the model can be used in such process.

We rely on the downscaled annual burn probabilities, provided in Section 4.5.1, to be the measure of hazard for each geographical location. With no access to vulnerability, we assume complete destruction of buildings and infrastructure where a large wildfire occurs, which will likely overestimate premiums. We choose to use the burn probabilities of large wildfire because insurance claims are more likely to occur from wildfires that are large in size, which cause complete destruction of the area. Hence, by using the CatIQ industry exposure database, the annual pure premium for wildfire insurance can be computed as the product of the annual burn probabilities of large wildfire by the cost of household replacement.

As an illustration of the technique, we analyze the distribution of pure premiums over British Colubmia. The predicted insurance pure premium to cover wildfire risk in British Columbia is 0 for around 89% of the high resolution grid cells. This is because those cells have predicted annual burn probability of 0. Table 4.10 summarizes the quantiles of the insurance pure premium per household over all grid cells.

Table 4.10: Summary statistics of the distribution of the predicted insurance pure premium per household in British Columbia.

Quantile level	90%	95%	97.5%	99%	99.5%	100%
Pure premium	30	59	208	1,147	1,753	7,980

4.6 Discussion

Wildfire risk in North America is of great significance and over the past decade, insurers have been suffering annually from billions of dollars in losses. This article proposes a transparent and simple, yet powerful model that relies on random forests. It can be used by actuaries to price, reserve or manage the financial risk from wildfires. Unlike wildfire

models that were designed for a specific region, the models presented in this article cover a wide geographic area. They have high predictive capabilities with strong ability to classify locations into high/low risk zones. The random forest models are characterised by their ability to capture non-linearity and interaction in the model inputs. Occurrence of wildfire can be devastating, thus it is of utmost importance to measure wildfire risk by calculating the burn probabilities of wildfires in high-risk areas with densely populated communities. With sufficient property characteristics data obtained at the underwriting stage, our model can be used to compute the loss costs necessary for premium calculation, accordingly, complementing vendor catastrophe-loss models.

Calculating insurance premiums is a very important task for actuaries, and it is a complex endeavor for emerging risks such as wildfires. The approach described in Section 5 along with the Leaflet maps provided (see Supplementary Material) can be used as a baseline description of the wildfire hazard. With extensive property characteristics data obtained at the underwriting stage and appropriate vulnerability curves, actuaries can therefore compute the loss costs necessary for premium determination and accurate underwriting. Moreover, some actuarial applications such as reserving or capital requirements might require a representation of spatial dependence. By construction, the wildfire models presented in Section 4.3 link occurrence to a set of covariates, and as such, borrows spatial dependence from e.g., temperature and precipitation dynamics. The actuary shall therefore make sure the inputs to the wildfire model are spatially consistent over the desired time horizon for the aforementioned applications.

Supplementary Material

Leaflet maps for the annual burn probabilities of wildfires of any size and of large wildfires are available in https://robabairakdar.shinyapps.io/WildfireLikelihood/, and

https://robabairakdar.shinyapps.io/LargeWildfireLikelihood/, respectively. The maps are available for the three high-risk states and provinces: Alberta, British Columbia and California. They are desktop and mobile friendly interactive maps that provide the user with the flexibility to zoom on a certain location on a map. The leaflets can be used by insurers to obtain the burn probabilities for any location in the selected high-risk states and provinces.

Chapter 5

Conclusion

This Chapter concludes the thesis. Three manuscripts that focus on catastrophic risks and extreme events were provided.

In the first manuscript, we focus on the generalized tail distortion risk measure for extreme risks, which is used to assess tail risks of excess losses modeled by the right tails of loss distributions. We explore its asymptotic expansions, and rely on them to provide an estimator for this risk measure. We prove the asymptotic normality of the estimator at intermediate and extreme confidence levels. We also provide bias-corrected estimators that use the Hill estimator to estimate the tail index. Based on the results of the simulation study, we can confirm that our estimators provide good results, especially for larger sample sizes. The application section showcases the usage of the estimators on real-life actuarial data sets.

In the second manuscript, we extend the methods available in the literature to assess the goodness-of-fit of spatio-temporal models through residual analysis methods. Specifically, we define and implement deviance Voronoi residuals, which provide a very useful tool to compare fitted models and identify locations where one model is superior to the other. We also create an earthquake financial losses estimation tool for Canada by relying on

building information, their replacement costs and earthquake damage probability matrices. Additionally, insurance policy terms and market information are used to estimate insurance claim values. A more interpretable approach is suggested to calculate the county-wide PML by relying on the correlation between neighboring provinces. A large simulation of earthquakes is performed which concluded that some Canadian provinces are ill-prepared for earthquake relief and that there is a need for further insurance market penetration, especially in Eastern Canada. Our simulation-based approach can inform policymakers on disaster resilience and help the public improve their risk awareness and perhaps increase insurance uptake. We also provide an open-source interactive web application that allows the users to simulate multiple significant earthquakes, with different random moment magnitudes, in a chosen geographical location in Canada. The insurance policy terms can be set by the user to calculate insurance claims. This web application can provide insurers with a simulated value of the expected financial losses in case of an occurrence of a significant earthquake in areas where they have exposure or plan to sell new earthquake policies.

In the third manuscript, we propose a transparent and simple, yet powerful model that utilizes random forest regression to model the occurrences of wildfires in North America. Occurrence of wildfire can be devastating, thus it is of utmost importance to measure wildfire risk by calculating the burn probabilities of wildfires in high-risk areas with densely populated communities. Unlike wildfire models that were designed for a specific region, the models presented in this article cover a wide geographic area. They have high predictive capabilities with strong ability to classify locations into high/low risk zones. Calculating insurance premiums is a very important task for actuaries, and it is a complex endeavor for emerging risks such as wildfires. The catastrophe risk decomposition approach described along with the Leaflet maps provided can be used as a baseline description of the wildfire hazard. With extensive property characteristics data obtained at the underwriting stage and appropriate vulnerability curves, actuaries can therefore compute the loss costs

necessary for premium determination and accurate underwriting. Additionally, the model can provide regulators with a more uniform and transparent tool with which to assess insurer solvency risk. Moreover, some actuarial applications such as reserving or capital requirements might require a representation of spatial dependence. By construction, the wildfire models presented in Chapter 4 link occurrence to a set of covariates, and as such, borrows spatial dependence from e.g., temperature and precipitation dynamics. The actuary shall therefore make sure the inputs to the wildfire model are spatially consistent over the desired time horizon for the aforementioned applications. Possible future research includes modeling the size of the wildfire. This can be performed by using propagation models that rely on meteorological and topographical information to explain the direction and speed of spread of wildfires. It is also important to consider fire control measures and proximity to water.

Chapter 6

Appendices

This appendix supplements the chapters of this thesis. This includes additional details, definitions, algorithms and results that are mentioned within the text.

I Deviance Voronoi Residuals for Space-time Point Process Models: An Application to Earthquake Insurance Risk

I.1 Modified Mercalli Intensity Definitions

Table 6.1: Modified Mercalli Intensity Definitions (Wood and Neumann 1931)

Intensity	**Shaking**	**Description**
I	Not felt	Not felt except by very few under especially favorable conditions.
II	Weak	Felt only by a few people at rest, especially on upper floors of buildings. Delicately suspended objects may swing.

III	Weak	Felt quite noticeably by people indoors, especially on upper floors of buildings: Many people do not recognize it as an earthquake. Standing motor cars may rock slightly. Vibrations are similar to the passing of a truck.
IV	Light	Felt indoors by many, outdoors by few during the day: At night, some are awakened. Dishes, windows, and doors are disturbed; walls make cracking sounds. Sensations are like a heavy truck striking a building. Standing motor cars are rocked noticeably.
V	Moderate	Felt by nearly everyone; many awakened: Some dishes and windows are broken. Unstable objects are overturned.
VI	Strong	Felt by all, and many are frightened. Some heavy furniture is moved; a few instances of fallen plaster occur. Damage is slight.
VII	Very strong	Damage is negligible in buildings of good design and construction; but slight to moderate in well-built ordinary structures; damage is considerable in poorly built or badly designed structures; some chimneys are broken. Noticed by people in driving motor cars.
VIII	Severe	Damage slight in specially designed structures; considerable damage in ordinary substantial buildings with partial collapse. Damage great in poorly built structures. Fall of chimneys, factory stacks, columns, monuments, walls. Heavy furniture overturned. Sand and mud ejected in small amounts. Changes in well water. People in driving motor cars are disturbed.

IX	Violent	Damage is considerable in specially designed structures; well-designed frame structures are thrown out of plumb. Damage is great in substantial buildings, with partial collapse. Buildings are shifted off foundations. Liquefaction occurs. Underground pipes are broken.
X	Extreme	Some well-built wooden structures are destroyed; most masonry and frame structures are destroyed with foundations. Rails are bent. Landslides considerable from river banks and steep slopes. Shifted sand and mud. Water splashed over banks.
XI	Extreme	Few, if any, (masonry) structures remain standing. Bridges are destroyed. Broad fissures erupt in the ground. Underground pipelines are rendered completely out of service. Earth slumps and land slips in soft ground. Rails are bent greatly.
XII	Extreme	Damage is total. Waves are seen on ground surfaces. Lines of sight and level are distorted. Objects are thrown upward into the air.

I.2 Collection of building inventory and calculation of exposure

We define:

- Total square footage = # units in a building type × average square footage of that type of unit
- Building exposure = Total square footage × Mean replacement cost
- Building content exposure = Building exposure × % contents value.

Now we explain the sources of the building inventory data as well as the repair and

replacement cost information.

Residential Dwellings

Information on the number of buildings for each residential building classification is obtained from Statistics Canada, Government of Canada (2016). The building classifications (single-detached houses, row houses, etc.) are transformed into HAZUS' occupancy codes as per Table C.4 in Ulmi et al. (2014). The Canadian Housing Statistics Program contains comprehensive data on the average total living area in square feet for each residential building class in each CSD (Statistics Canada, Government of Canada 2020a). The data are currently available for BC, ON, NS and NB. For CSDs that have missing square footage values, we use the square footage of the CSD in the same province that has the closest household median income value, obtained from Statistics Canada, Government of Canada (2016). For CSDs that have missing square footage values and missing household median income value, we use the weighted average square footage of the province, where the weights are the number of houses in each CSD. For the provinces that do not have square footage values, we use the values of a CSD that has the nearest household income values, where the provinces are matched as follows:

- ON square footage is used for for QC, MB, and NU.
- NB square footage is used for NL and PE.
- BC square footage is used for SK, AB, YT, and NT.

For dwelling types that do not have square footage information, we assume that an "other single-attached" house has the same square footage as a "single-detached" house and an "apartment or flat in a duplex" has the same square footage as a "Semi-detached" house. We use the values suggested in Ulmi et al. (2014) for the remaining dwelling types that do not have information on square footage.

The building replacement cost, in dollar units, for each HAZUS occupancy code are obtained from HAZUS Ulmi et al. (2014), which was originally obtained from the RSMeans. The building construction price index is used to inflate the construction costs to June 2021 (Statistics Canada, Government of Canada 2020b). The BCPI is available for eleven census metropolitan areas, thus we generalize the inflation rate to each area's respective province. Some provinces/territories were not included in the data and thus we assumed that NL, NU, NT, YT, PE, SK and MB will follow the smallest inflation value, which is that of Edmonton, AB. We also assumed that NB will follow the inflation trend of Halifax, NS.

The building content replacement value for residential dwellings is assumed to be at 50% of the building replacement cost as suggested in Ulmi et al. (2014) and FEMA (2013). The building exposure for each HAZUS code is converted to construction types (wood, concrete, steel, masonry, etc.) by using HAZUS' general building scheme mapping information, available in Table 5.1 in FEMA (2013).

Non-residential Dwellings

Statistics Canada currently does not have a comprehensive dataset for non-residential buildings, thus, we will rely on their building permits data, available in Statistics Canada, Government of Canada (2020c). The average annual ratios of institutional and governmental, commercial and industrial building permits to residential building permits are calculated over the years 2011 to 2019. Accordingly, the exposure of non-residential buildings is calculated as a percentage of the residential buildings total exposure. Additionally, a detailed split of each building category is available in Statistics Canada, Government of Canada (2020c). Table 6.2 provides the suggested conversion of non-residential building types to HAZUS occupancy codes.

Similar to residential buildings, the exposure by HAZUS dwelling type is converted to construction types (wood, concrete, steel, masonry, etc.), following Table 5.1 in FEMA (2013). The building content replacement value for non-residential dwellings is assumed to

be a percentage of the building replacement cost, as suggested by FEMA (2013).

Table 6.2: Conversion from Statistics Canada classification to HAZUS occupancy codes for non-residential buildings

Statistics Canada Building Permits Label	**HAZUS Occupancy Code**	**Description**
Institutional and governmental		
Elementary school, kindergarten	EDU1	Grade Schools
Secondary school, high school, junior high school	EDU1	Grade Schools
Post-secondary institution and technical institute	EDU2	Colleges/Universities
University	EDU2	Colleges/Universities
Library, museum, art gallery, aquarium, botanical garden, scientific center	COM8	Entertainment, recreation
General hospital	COM6	Hospital
Clinic, out-patient clinic, first aid station	COM7	Medical Office/Clinic
Welfare, home	RES5	Institutional dormitory
Churches, religion	REL1	Church/Non-Profit
Government legislative and administration building, city hall, court of justice, embassy, parliament and senate building	GOV1	General services
Other government building - police station, prison, fire station, military building	GOV2	Emergency response

Industrial		
Maintenance building	IND1	Heavy factory
Plant for manufacturing, processing and assembling goods	IND1	Heavy factory
Communication building	IND1	Heavy factory
Transportation terminal	IND1	Heavy factory
Utility building	IND1	Heavy factory
Mining building	IND4	Metals/minerals processing
Agriculture	AGR1	Agriculture
Commercial		
Trade and services	COM1	Retail trade
Warehouses	COM2	Wholesale trade
Service stations	COM3	Personal and repair services
Office buildings	COM4	Professional/technical services
Theatre and performing art center, movie theatre, concert hall, opera house, cultural center	COM9	Theaters
Indoor recreational building, sports complex, tennis court and squash, community center, arena, curling club, swimming pool	COM8	Entertainment, recreation
Outdoor recreational building, country club, golf club campground facilities, outdoor skating rink, outdoor swimming pool	COM8	Entertainment, recreation

Convention center, exhibition building	COM4	Professional/technical services
Hotel, hotel and motel, motor hotel	RES4	Temporary lodging
Motel, cabin for tourism	RES4	Temporary lodging
Student's residence, boarding house, religious residence, hostel, dormitory	RES5	Institutional dormitory
Restaurant, bar, night club, diner	COM8	Entertainment, recreation
Laboratories	COM7	Medical Office/Clinic

I.3 Residual Analysis of the Fitted STPP Models

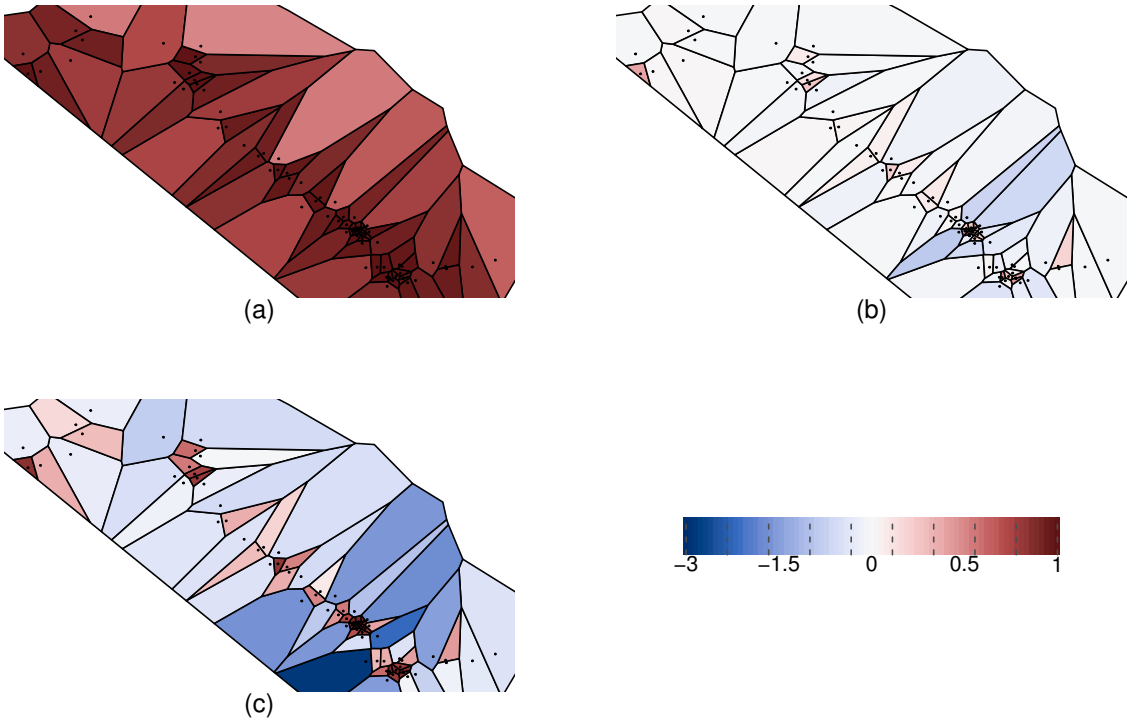


Figure 6.1: Focusing on Western Canada, raw Voronoi residuals of models P (a), H_1 (b), H_2 (c).

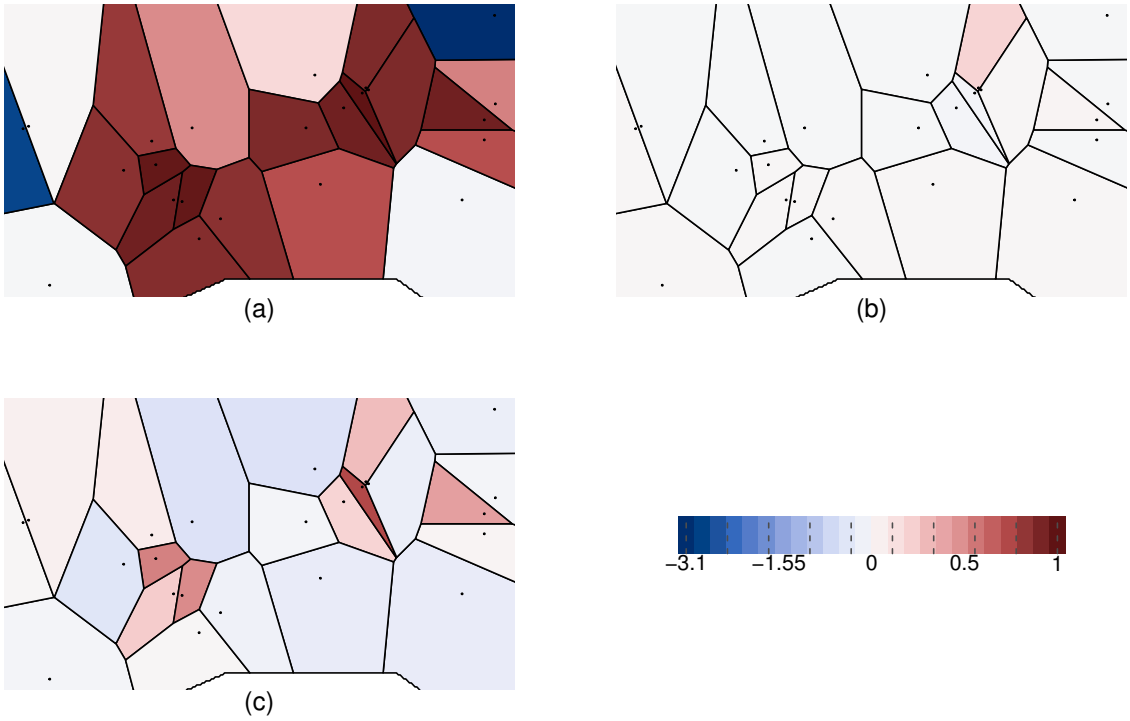


Figure 6.2: Focusing on Eastern Canada, raw Voronoi residuals of models P (a), H_1 (b), and H_2 (c).

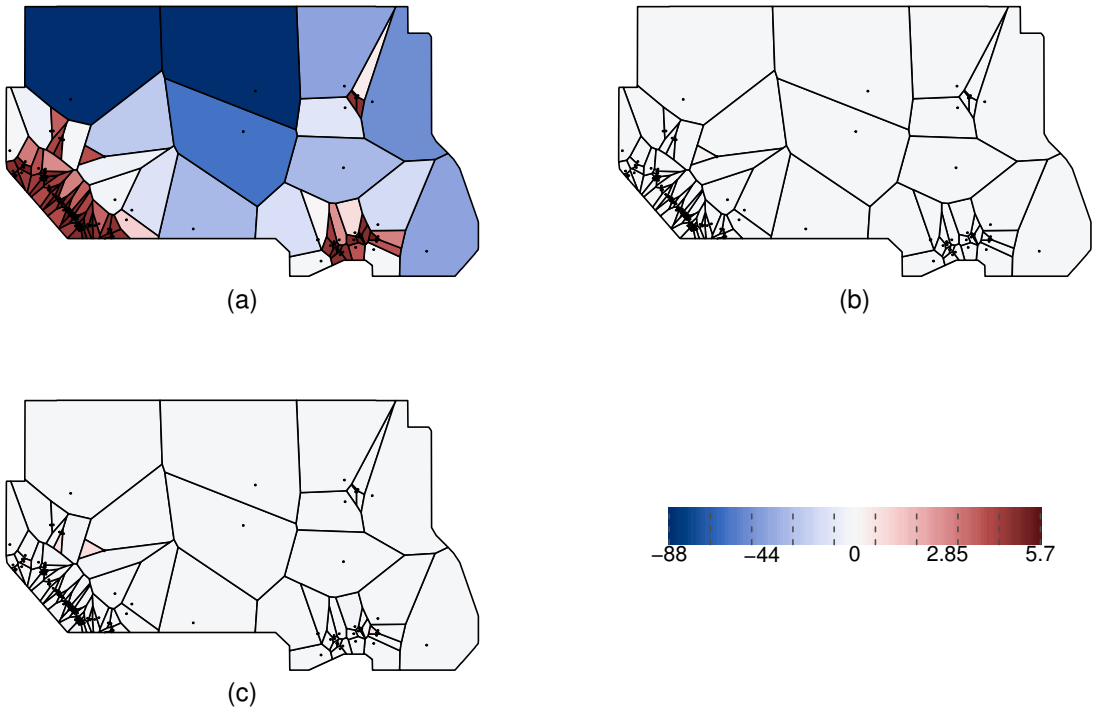


Figure 6.3: Pearson Voronoi residuals of models P (a), H_1 (b), and H_2 (c).

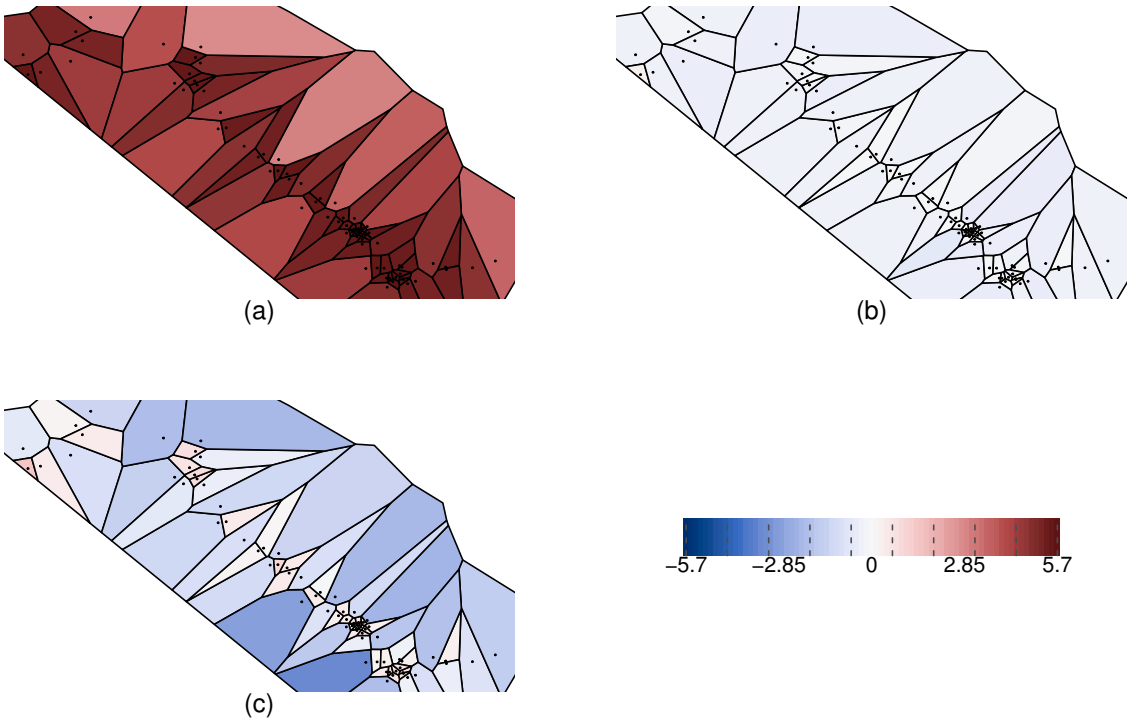


Figure 6.4: Focusing on Western Canada, Pearson Voronoi residuals of models P (a), H_1 (b), and H_2 (c).

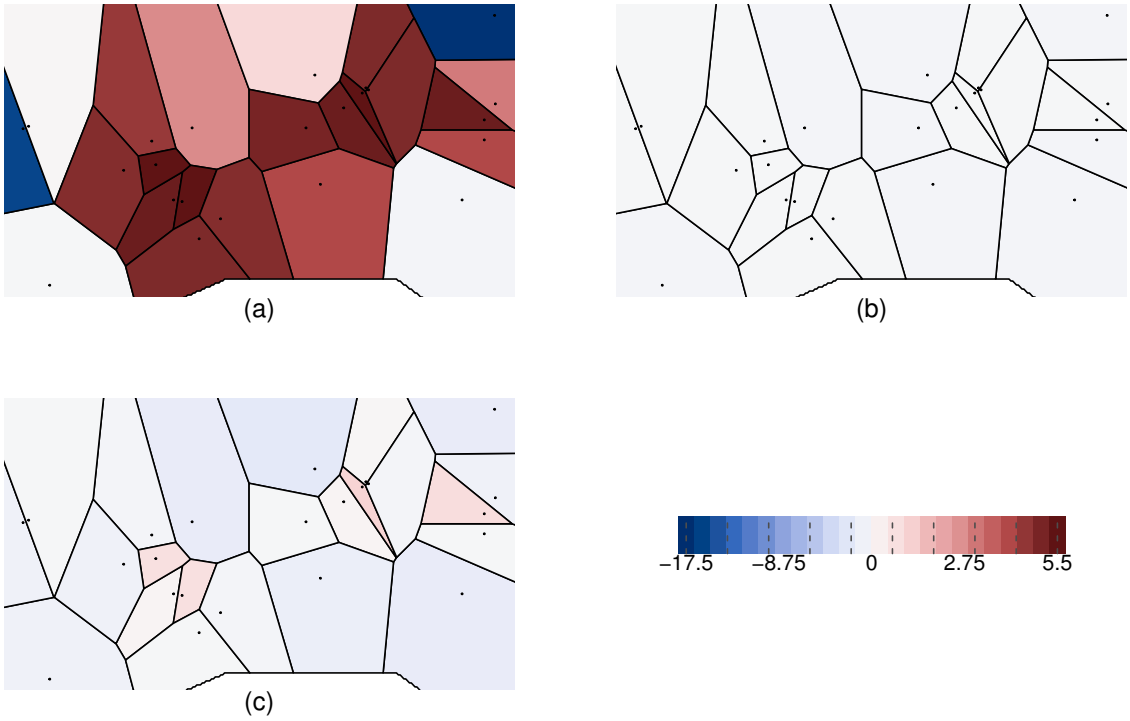


Figure 6.5: Focusing on Eastern Canada, Pearson Voronoi residuals of models P (a), H_1 (b), and H_2 (c).

I.4 Extreme Value Theory

Extreme value analysis is a branch in statistics that is focused on the behavior of the tail of the distribution (Coles 2001). There are two principle models for extreme values: the block maxima model and the peaks-over-threshold model. The *block maxima* approach is used to model the largest observations from samples of identically distributed observations in successive blocks. The *peaks-over-threshold* is used to model all large observations that exceed a given high threshold value, denoted u.

Distribution Of The Maxima

The limiting distribution of block maxima is given in the following theorem:

Theorem I.1. ***Fisher and Tippet theorem*** *(Fisher and Tippett 1928)*
Let $X_1, \ldots, X_n$ be a sequence of independent random variables having a common distribution function F and consider $M_n = \max\{X_1, \ldots, X_n\}$. If there exists norming constants (a_n) and (b_n), where $a_n \in \mathbb{R}$ and $b_n > 0$ for all $n \in \mathbb{N}$ and some non-degenerate distribution function H such that

$$\frac{M_n - a_n}{b_n} \xrightarrow{d} H,$$

then H belongs to one of the following three classes of distributions (up to location and scaling):

$$\textit{Fréchet:} \quad \Phi_\alpha(x) = \begin{cases} 0, & x \leq 0, \\ \exp\left\{-x^{-\alpha}\right\}, & x > 0, \end{cases} \qquad \alpha > 0$$

$$\textit{Gumbel:} \quad \Lambda(x) = \exp\left\{-\exp\{-x\}\right\}, \quad x \in \mathbb{R},$$

$$\textit{Weibull:} \quad \Psi_\alpha(x) = \begin{cases} \exp\left\{-(-x)^{\alpha}\right\}, & x \leq 0, \\ 1, & x > 0, \end{cases} \qquad \alpha > 0$$

Results by Von Mises (1954) and Jenkinson (1955) provide a generalization of Theorem I.1. Set $\xi = \alpha^{-1}$ for the Fréchet distribution, $\xi = -\alpha^{-1}$ for the Weibull distribution and interpret the Gumbel distribution as a limiting case as $\xi \to 0$, then we obtain the following definition.

Definition I.1. ***Generalized Extreme Value Distribution.***

The distribution function of a GEV is given by

$$H_\xi(x) = \begin{cases} \exp\left\{-\left(1 + \xi x\right)^{-\frac{1}{\xi}}\right\}, & \xi \neq 0, \\ \exp\left\{-\exp(-x)\right\}, & \xi = 0, \end{cases}$$

where $1 + \xi x > 0$. A three-parameter family is obtained by defining $H_{\xi,\mu,\sigma} := H_\xi\left(\frac{x-\mu}{\sigma}\right)$ for

a location parameter $\mu \in \mathbb{R}$, a scale parameter $\sigma > 0$, and a shape parameter $\xi \in \mathbb{R}$.

Distribution Of The Exceedances

Instead of considering only the maximum from a block, the peaks-over-threshold method rather considers all the exceedances over some high threshold value u.

Definition I.2. ***Excess Distribution over threshold u.***
Let X be a random variable with distribution function F and an upper end-point $x_F \leq \infty$, then the excess distribution over the threshold u is defined as

$$F_u(x) = Pr(X - u \leq x | X > u) = \frac{F(x+u) - F(u)}{1 - F(u)}, \quad 0 \leq x < x_F - u.$$

Theorem I.2. *(Pickands III 1975; Balkema and De Haan 1974)*
If F is a distribution function that belongs to the maximum domain attraction of a GEV distribution $H_{\xi,\mu,\sigma}$, i.e. if F satistifies the conditions of Theorem I.1, then

$$\lim_{u \to x_F} \sup_{0 \leq x < x_F - u} |F_u(x) - G_{\xi,\sigma}(x)| = 0,$$

where

$$G_{\xi,\sigma}(x) = \begin{cases} 1 - \left(1 + \xi \frac{x}{\sigma}\right)^{-\frac{1}{\xi}}, & \xi \neq 0, \\ 1 - \exp\left(-\frac{x}{\sigma}\right), & \xi = 0, \end{cases}$$

for $\sigma > 0$, and $x \geq 0$ when $\xi \geq 0$, while $0 \leq x \leq -\sigma/\xi$ when $\xi < 0$. The parameters ξ and σ are referred to, respectively, as the shape and scale parameters.

We assume $F_u \approx G_{\xi,\sigma}$ for a high threshold u, where $G_{\xi,\sigma}$ is called the Generalized Pareto Distribution (GPD).

A large quantile is defined as the return level x_m that is exceeded on average once every

m observations provided that m is large enough and satisfying $x_m > u$, where

$$\Pr(X > x_m) = \Pr(X > u) \cdot \left[1 + \xi \left(\frac{x_m - u}{\sigma}\right)\right]^{-1/\xi} = \frac{1}{m}.$$

By solving for x_m,

$$x_m = u + \frac{\sigma}{\xi}\left[(m \cdot \Pr(X > u))^{\xi} - 1\right]. \tag{I.1}$$

Poisson Approximation of the Number of Excesses over a High Threshold

Theorem I.3. *Let $X_1, \ldots, X_n$ be a sequence of independent random variables satisfying the conditions in Theorem I.2, and let N_n be the number of excesses over a threshold u_n. If the sequence of threshold (u_n) satisfies*

$$\lim_{n\to\infty} n(1 - F(u_n)) = \lambda,$$

then, for $k = 0, 1, 2, \ldots,$

$$\lim_{n\to\infty} \mathbb{P}(N_n \leq k) = \sum_{s=0}^{k} \frac{e^{-\lambda}\lambda^s}{s!}.$$

Thus, assuming a high threshold u, the number of exceedances can be modelled as a marked homogeneous Poisson process with rate λ and the size of the excesses is the limiting GPD. Accordingly, the point process model requires the estimation of the rate λ of the homogeneous Poisson process and the parameters of the GPD.

Distribution Of The Maximum Of Homogeneous Poisson Process With GPD Marks

Let N be a Poisson random variable with mean λ and let $X_1, \ldots, X_N$ be a sequence of N independent and identically distributed random variables with common cumulative distribution function $G_{\xi,\sigma}$. Let $M_N = \max\{X_1, \ldots, X_N\}$, then Cebrián, Denuit, and

Lambert (2003) indicates that

$$\mathbb{P}(M_N \leq x) = H_{\xi,\mu,\psi}(x),$$

with $\mu = \sigma/\xi \left(\lambda^\xi - 1\right)$ and $\psi = \sigma\lambda^\xi$.

I.5 Algorithms

Algorithm 1: Earthquake losses and claims simulation

Result: $\text{PML}_{1/\epsilon}$ of Eastern and Western Canada earthquake losses and claims

$n \leftarrow 100000$

SpatioTemporalSim$\leftarrow$ n years of simulated earthquakes from a spatio-temporal point process

for $i \leftarrow 1$ **to** length(**SpatioTemporalSim**) **do**

  SimEQ$_i \leftarrow$ coordinates of the i^{th} earthquake in **SpatioTemporalSim**

  $\text{E}_i \leftarrow \mathbb{1}_{(\text{Longitude of } \textbf{SimEQ}_i > -100)}$

  Grid$_i \leftarrow$ coordinates of the nearest neighbor from the PGA Grid to **SimEQ**$_i$

  $\text{PGA}_i \leftarrow$ a simulated PGA value from a fitted GPD, whose parameters are estimated from the 8 PGA quantiles in **Grid**$_i$. PGA_i must be corresponding to a moment magnitude > 6.

  $\text{MMI}_i \leftarrow 3.66 \log(\text{PGA}_i) - 1.66$, for PGA_i in cm/s^2

  $\text{d}_i \leftarrow$ distance in km between **SimEQ**$_i$ and **Grid**$_i$.

  if $\text{E}_i == 1$ **then**

    $\text{Magnitude}_i = (\text{MMI}_i - 1.41 + 2.08 \log_{10}(\text{d}_i) + 0.00345\text{d}_i)\,/1.68,$

  else

    $\text{Magnitude}_i = (\text{MMI}_i - 5.07 + 3.69 \log_{10}(\text{d}_i))/1.09,$

  RadiusMMI$_i \leftarrow$ the radii of MMI circles centered at **SimEQ**$_i$ for MMI levels from MMI_i to VI, for a given value Magnitude_i.

  for $j \leftarrow 1$ **to** length(**RadiusMMI**$_i$) **do**

    CSD$_{i,j} \leftarrow$ list of CSDs that intersect MMI circle of radius **RadiusMMI**$_{i,j}$

    CSDAreaPrcnt$_{i,j} \leftarrow$ the percentage of land of each CSD that intersects the MMI circle of radius **RadiusMMI**$_{i,j}$

    `CalculateLossesandClaims` (MMI_i-[j-1],**CSD**$_{i,j}$, **CSDAreaPrcnt**$_{i,j}$)

Algorithm 2: Calculate Losses and Claims Function

Input : MMI of earthquake (MMI), list of CSDs (**CSD**), percentage affected for each CSD (**CSDAreaPrcnt**)

Output: Losses and Claims

foreach CSD $C \in$ **CSD do**

All steps are done for **CSD** C, but the subscript is removed for readability.

Let $\text{Loss}_{t,B}$ be the losses for building type B and damage type t, for $t \in \{S, AS, DS, BldgC\}$.

Let $\text{Exposure}_{t,B}$ be the building and/or building contents exposure for building type B and damage type t, for $t \in \{S, AS, DS, BldgC\}$.

Let $\text{DamageRange}_{t,B,i}$ be the damage range associated with damage state i, building type B, and damage type t, for $t \in \{S, AS, DS, BldgC\}$ for a given MMI level. Let $\text{pDamageRange}_{t,B,i}$ be its associated probability.

Let $\text{MDF}_{t,B,i}$ be the mean damage factor associated with damage state i for building type B and damage type t, for $t \in \{S, AS, DS, BldgC\}$.

Let DedPrcnt be the percentage deductible

Let LmtPrcnt be the percentage policy limit

Let InsPrcnt be the percentage of insurance market penetration

foreach Building type B **do**

Let BldgCostUncertainty be the uncertainty associated to the cost of building replacement.

$\text{BldgCostUncertainty} \leftarrow \mathcal{U}(0.9, 1.1)$

$\text{Ded}_B \leftarrow$ **CSDAreaPrcnt** $\times$ DedPrcnt $\times \sum_t \text{Exposure}_{t,B} \times$ BldgCostUncertainty

$\text{Lmt}_B \leftarrow$ **CSDAreaPrcnt** $\times$ LmtPrcnt $\times \sum_t \text{Exposure}_{t,B} \times$ BldgCostUncertainty

foreach Damage type $t \in \{\text{S, AS, DS, BldgC}\}$ **do**

$\text{MDF}_{t,B,i} \leftarrow \left[\text{sample from } \mathcal{U}\left(\text{DamageRange}_{t,B,i}\right)\right] \times \text{pDamageRange}_{t,B,i}$

$\text{MDF}_{t,B} \leftarrow \sum_i \text{MDF}_{t,B,i}$

$\text{Loss}_{t,B} \leftarrow \text{Exposure}_{t,B} \times \text{BldgCostUncertainty} \times \text{MDF}_{t,B} \times \textbf{CSDAreaPrcnt}$

$\text{LOSS}_B \leftarrow \sum_t \text{Loss}_{t,B}$

$\text{CLAIM}_B \leftarrow \text{InsPrcnt} \times \max\{0, \min\{\text{LOSS}_B - \text{Ded}_B, \text{Lmt}_B - \text{Ded}_B\}\}$

$\text{LOSS} = \sum_B \text{LOSS}_B$

$\text{CLAIM} = \sum_B \text{CLAIM}_B$

return $\sum_C \text{LOSS}, \sum_C \text{CLAIM}$

I.6 Correlation of losses and claims

Table 6.3: Pearson correlation coefficient of the simulated insurance claims between Canadian provinces, based on 100,000 years of simulated earthquakes.

	NL	**PE**	**NS**	**NB**	**QC**	**ON**	**MB**	**SK**	**BC**	**YT**	**NT**	**AB**	**NU**
NL	1.00	0.29	0.23	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PE	0.29	1.00	0.77	0.89	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NS	0.23	0.77	1.00	0.87	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NB	0.23	0.89	0.87	1.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
QC	0.00	0.00	0.00	0.02	1.00	0.69	0.48	0.00	0.00	0.00	0.00	0.00	0.00
ON	0.00	0.00	0.00	0.00	0.69	1.00	0.60	0.00	0.00	0.00	0.00	0.00	0.00
MB	0.00	0.00	0.00	0.00	0.48	0.60	1.00	0.01	0.00	0.00	0.03	0.00	0.03
SK	0.00	0.00	0.00	0.00	0.00	0.00	0.01	1.00	0.01	0.02	0.09	0.01	0.07
BC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	1.00	0.16	0.08	0.63	0.11
YT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.16	1.00	0.39	0.06	0.34
NT	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.09	0.08	0.39	1.00	0.04	0.82
AB	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.63	0.06	0.04	1.00	0.07
NU	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.07	0.11	0.34	0.82	0.07	1.00

Table 6.4: Kendall's tau of the simulated financial losses between Canadian provinces, based on 100,000 years of simulated earthquakes.

	NL	**PE**	**NS**	**NB**	**QC**	**ON**	**MB**	**SK**	**BC**	**YT**	**NT**	**AB**	**NU**
NL	1.00	0.73	0.75	0.39	0.22	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PE	0.73	1.00	0.78	0.52	0.33	0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NS	0.75	0.78	1.00	0.53	0.33	0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NB	0.39	0.52	0.53	1.00	0.74	0.65	0.19	0.00	0.00	0.00	0.00	0.00	0.00
QC	0.22	0.33	0.33	0.74	1.00	0.88	0.43	0.00	0.00	0.00	0.00	0.00	0.00
ON	0.09	0.19	0.19	0.65	0.88	1.00	0.51	0.01	0.00	0.00	0.00	0.00	0.00
MB	0.00	0.00	0.00	0.19	0.43	0.51	1.00	0.30	0.02	0.27	0.27	0.09	0.30
SK	0.00	0.00	0.00	0.00	0.00	0.01	0.30	1.00	0.20	0.57	0.61	0.44	0.43
BC	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.20	1.00	0.34	0.27	0.39	0.36
YT	0.00	0.00	0.00	0.00	0.00	0.00	0.27	0.57	0.34	1.00	0.79	0.55	0.65
NT	0.00	0.00	0.00	0.00	0.00	0.00	0.27	0.61	0.27	0.79	1.00	0.45	0.70
AB	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.44	0.39	0.55	0.45	1.00	0.36
NU	0.00	0.00	0.00	0.00	0.00	0.00	0.30	0.43	0.36	0.65	0.70	0.36	1.00

Table 6.5: Kendall's tau of the simulated insurance claims between Canadian provinces, based on 100,000 years of simulated earthquakes.

	NL	**PE**	**NS**	**NB**	**QC**	**ON**	**MB**	**SK**	**BC**	**YT**	**NT**	**AB**	**NU**
NL	1.00	0.69	0.70	0.36	0.21	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PE	0.69	1.00	0.82	0.51	0.33	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NS	0.70	0.82	1.00	0.52	0.34	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NB	0.36	0.51	0.52	1.00	0.75	0.65	0.20	0.00	0.00	0.00	0.00	0.00	0.00
QC	0.21	0.33	0.34	0.75	1.00	0.87	0.46	0.00	0.00	0.00	0.00	0.00	0.00
ON	0.08	0.17	0.17	0.65	0.87	1.00	0.55	0.04	0.00	0.00	0.00	0.00	0.00
MB	0.00	0.00	0.00	0.20	0.46	0.55	1.00	0.22	0.00	0.15	0.15	0.00	0.19
SK	0.00	0.00	0.00	0.00	0.00	0.04	0.22	1.00	0.11	0.36	0.50	0.21	0.30
BC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	1.00	0.27	0.21	0.33	0.30
YT	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.36	0.27	1.00	0.71	0.34	0.57
NT	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.50	0.21	0.71	1.00	0.23	0.59
AB	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.21	0.33	0.34	0.23	1.00	0.23
NU	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.30	0.30	0.57	0.59	0.23	1.00

Bibliography

AIR Worldwide. 2013. Study of Impact and the Insurance and Economic Cost of a Major Earthquake in British Columbia and Ontario/Québec. *Insurance Bureau of Canada* 345.

Anderson, D., S. Feldblum, C. Modlin, D. Schirmacher, E. Schirmacher, and N. Thandi. 2004. A Practitioner's Guide to Generalized Linear Models. *Casualty Actuarial Society Discussion Paper Program,* 1–116.

Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath. 1999. Coherent measures of risk. *Mathematical finance* 9 (3): 203–228.

ATC (Applied Technology Council, Redwood City, California). 1985. *ATC-13: Earthquake Damage Evaluation for California.*

Atkinson, D., M. Chladil, V. Janssen, and A. Lucieer. 2010. Implementation of Quantitative Bushfire Risk Enalysis in a GIS environment. *International Journal of Wildland Fire* 19 (5): 649–658.

Baddeley, A., J. Møller, and A. G. Pakes. 2008. Properties of residuals for spatial point processes. *Annals of the Institute of Statistical Mathematics* 60 (3): 627–649.

Baddeley, A., E. Rubak, and R. Turner. 2015. *Spatial Point Patterns: Methodology and Applications with R.* CRC Press.

Baddeley, A., and R. Turner. 2005. spatstat: An R Package for Analyzing Spatial Point Patterns. *Journal of Statistical Software* 12 (6): 1–42. https://doi.org/10.18637/jss.v012.i06.

Baddeley, A., R. Turner, J. Møller, and M. Hazelton. 2005. Residual Analysis for Spatial Point Processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (5): 617–666.

Bakun, W. H., A. C. Johnston, and M. G. Hopper. 2003. Estimating Locations and Magnitudes of Earthquakes in Eastern North America from Modified Mercalli Intensities. *Bulletin of the Seismological Society of America* 93 (1): 190–202.

Bakun, W. H., and C. M. Wentworth. 1997. Estimating Earthquake Location and Magnitude from Seismic Intensity Data. *Bulletin of the Seismological Society of America* 87 (6): 1502–1521.

Balkema, A. A., and L. De Haan. 1974. Residual Life Time at Great Age. *The Annals of Probability* 2 (5): 792–804.

Beirlant, J., Y. Goegebeur, J. Segers, and J. L. Teugels. 2004. *Statistics of extremes: theory and applications.* Vol. 558. John Wiley & Sons.

Boots, B., A. Okabe, and K. Sugihara. 1999. Spatial Tessellations. *Geographical Information Systems* 1:503–526.

Bray, A., K. Wong, C. D. Barr, and F. P. Schoenberg. 2014. Voronoi Residual Analysis of Spatial Point Process Models with Applications to California Earthquake Forecasts. *The Annals of Applied Statistics* 8 (4): 2247–2267.

Breiman, L. 2001. Random Forests. *Machine Learning* 45 (1): 5–32.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 2001. *Classification and Regression Trees.* Routledge.

Bruce, P., and A. Bruce. 2017. *Practical Statistics for Data Scientists: 50 Essential Concepts.* O'Reilly Media, Inc.

CALFIRE (California Fire Department of Forestry and Fire Protection). 2021. "California Fire Incidents Overview." Accessed November 18, 2021. https://www.fire.ca.gov/incidents/.

CALFIRE (California Fire Department of Forestry and Fire Protection). 2022. "Top 20 Largest California Wildfires." Accessed February 8, 2022. https://www.fire.ca.gov/.

CatIQ (Canada's Loss and Exposure Indices Provider). 2021. Accessed October 15, 2021. https://public.catiq.com.

CDS (Copernicus Climate Change Service (C3S) Climate Data Store). 2019. "ERA5-Land Hourly Data from 1950 to present." Accessed March 15, 2021.

Cebrián, A. C., M. Denuit, and P. Lambert. 2003. Generalized Pareto fit to the Society of Actuaries' Large Claims Database. *North American Actuarial Journal* 7 (3): 18–36.

Cecil, D. J., D. E. Buechler, and R. J. Blakeslee. 2014. Gridded Lightning Climatology from TRMM-LIS and OTD: Dataset Description. *Atmospheric Research* 135:404–414.

CFS (Canadian Forest Service). 2019. *National Fire Database - Agency Fire Data.* Natural Resources Canada, Canadian Forest Service, Northern Forestry Centre, Edmonton, Alberta. Accessed December 15, 2020. http://cwfis.cfs.nrcan.gc.ca/en_CA/nfdb.

Chuvieco, E., I. Aguado, M. Yebra, H. Nieto, J. Salas, M. P. Martín, L. Vilar, J. Martínez, S. Martín, P. Ibarra, et al. 2010. Development of a Framework for Fire Risk Assessment Using Remote Sensing and Geographic Information System Technologies. *Ecological Modelling* 221 (1): 46–58.

CIESIN (Center for International Earth Science Information Network) - Columbia University. 2017. "Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 11 (NASA Socioeconomic Data and Applications Center (SEDAC)." Accessed March 15, 2021. https://doi.org/10.7927/H49C6VHW.

CIFFC (Canadian Interagency Forest Fire Centre). 2021. "National Fire Situation Report, Sep 15, 2021." Accessed October 18, 2021. https://ciffc.net/en/ciffc/sitrep/2021-09-15.

Clements, R. A., F. P. Schoenberg, and A. Veen. 2012. Evaluation of Space-time Point Process Models Using Super-thinning. *Environmetrics* 23 (7): 606–616.

Coles, S. 2001. *An Introduction to Statistical Modeling of Extreme Values.* Vol. 208. Springer.

Cox, D. R., and V. Isham. 1980. *Point Processes.* Vol. 12. CRC Press.

Cragg, J. G. 1971. Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. *Econometrica: Journal of the Econometric Society,* 829–844.

Cressie, N. 2015. *Statistics for Spatial Data.* John Wiley & Sons.

Daley, D. J., and D. Vere-Jones. 2003. An Introduction to the Theory of Point Processes Volume 1: Elementary Theory and Methods. *Verlag New York Berlin Heidelberg: Springer.*

EM-DAT (Emergency Events Database). 2020. "EM-DAT: The International Disasters Database." Accessed November 15, 2021. http://www.emdat.be/.

De Jong, P., G. Z. Heller, et al. 2008. Generalized Linear Models for Insurance Data. *Cambridge Books.*

Denneberg, D. 1990. Distorted Probabilities and Insurance Premiums. *Methods of Operations Research* 63 (3): 3–5.

Denneberg, D. 1994. *Non-additive Measure and Integral.* Vol. 27. Springer Science & Business Media.

Denuit, M., and M. Mesfioui. 2017. Bounds on Kendall's Tau for Zero-inflated Continuous Variables. *Statistics & Probability Letters* 126:173–178.

Diggle, P. 1985. A Kernel Method for Smoothing Point Process Data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 34 (2): 138–147.

Diggle, P. 2013. *Statistical Analysis of Spatial and Spatio-temporal Point Patterns.* CRC Press.

El Methni, J., and G. Stupfler. 2017. Extreme versions of Wang risk measures and their estimation for heavy-tailed distributions. *Statistica Sinica,* 907–930.

El Methni, J., and G. Stupfler. 2018. Improved estimators of extreme Wang distortion risk measures for very heavy-tailed distributions. *Econometrics and Statistics* 6:129–148. https://doi.org/10.1016/j.ecosta.2017.03.002. https://doi.org/10.1016/j.ecosta.2017.03.002.

FEMA (Federal Emergency Management Agency). 2013. Multi-hazard Loss Estimation Methodology: Earthquake Model Hazus-MH 2.1 Technical Manual. *Department of Homeland Security, Federal Emergency Management Agency, Washington, DC.*

Field, E. H. 2007. Overview of the Working Group for the Development of Regional Earthquake Likelihood Models (RELM). *Seismological Research Letters* 78 (1): 7–16.

Finney, M. A., C. W. McHugh, I. C. Grenfell, K. L. Riley, and K. C. Short. 2011. A Simulation of Probabilistic Wildfire Risk Components for the Continental United States. *Stochastic Environmental Research and Risk Assessment* 25 (7): 973–1000.

Fisher, R. A., and L. H. C. Tippett. 1928. "Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample." In *Mathematical Proceedings of the Cambridge Philosophical Society,* 24:180–190. 2. Cambridge University Press.

Friedman, J., T. Hastie, R. Tibshirani, et al. 2001. *The elements of Statistical Learning.* Vol. 1. 10. Springer Series in Statistics New York.

Gabriel, E., P. J. Diggle, B. Rowlingson, and F. J. Rodriguez-Cortes. 2022. *stpp: Space-Time Point Pattern Simulation, Visualisation and Analysis.* R package version 2.0-6. https://CRAN.R-project.org/package=stpp.

Gabrielli, A., R. Richman, and M. V. Wüthrich. 2020. Neural Network Embedding of the Over-dispersed Poisson Reserving Model. *Scandinavian Actuarial Journal* 2020 (1): 1–29.

Ghorbani, M., N. Vafaei, J. Dvořák, and M. Myllymäki. 2021. Testing the first-order separability hypothesis for spatio-temporal point patterns. *Computational Statistics & Data Analysis* 161:107245.

Giglio, L., L. Boschetti, D. P. Roy, M. L. Humber, and C. O. Justice. 2018. The Collection 6 MODIS Burned Area Mapping Algorithm and Product. *Remote Sensing of Environment* 217:72–85.

Gordon, J. S., R. A. Clements, F. P. Schoenberg, and D. Schorlemmer. 2015. Voronoi Residuals and Other Residual Analyses Applied to CSEP Earthquake Forecasts. *Spatial Statistics* 14:133–150.

Gouvernement du Québec. 2021. *Québec is Resilient and Confident - Budget Plan.* Accessed February 9, 2022. http://www.budget.finances.gouv.qc.ca/budget/2021-2022/en/documents/BudgetPlan_2122.pdf.

Groom, N. 2015. "Wildfires Push Californians to High-risk Insurance Market." Accessed September 30, 2021. https://www.reuters.com/article/us-usa-wildfires-insurance-insight-idUSKCN0RL0AG20150921.

Guelman, L. 2012. Gradient Boosting Trees for Auto Insurance Loss Cost Modeling and Prediction. *Expert Systems with Applications* 39 (3): 3659–3667.

Haan, L. de, and A. Ferreira. 2006. *Extreme value theory: an introduction.* Vol. 21. Springer.

Haberman, S., and A. E. Renshaw. 1996. Generalized Linear Models and Actuarial Science. *Journal of the Royal Statistical Society: Series D (The Statistician)* 45 (4): 407–436.

Halchuk, S. C., J. E. Adams, and T. I. Allen. 2015. Fifth Generation Seismic Hazard Model for Canada: Grid Values of Mean Hazard to be used with the 2015 National Building Code of Canada. *Geological Survey of Canada, Open File 8285* 1. https://doi.org/10.4095/297378.

Hanes, C. C., X. Wang, P. Jain, M.-A. Parisien, J. M. Little, and M. D. Flannigan. 2019. Fire-regime Changes in Canada Over the Last Half Century. *Canadian Journal of Forest Research* 49 (3): 256–269.

Haouas, N., A. Necir, and B. Brahimi. 2019. Estimating the second-order parameter of regular variation and bias reduction in tail index estimation under random truncation. *Journal of Statistical Theory and Practice* 13 (1): 7.

Henckaerts, R., M.-P. Côté, K. Antonio, and R. Verbelen. 2021. Boosting Insights in Insurance Tariff Plans with Tree-based Machine Learning Methods. *North American Actuarial Journal* 25 (2): 255–285.

IBC (Insurance Bureau of Canada). 2021. "Lytton Wildfire Causes $78 Million in Insured Damage." Accessed October 15, 2021. http://www.ibc.ca/bc/resources/media-centre/media-releases/lytton-wildfire-causes-78-million-in-insured-damage.

Jenkinson, A. F. 1955. The Frequency Distribution of the Annual Maximum (or Minimum) values of Meteorological Elements. *Quarterly Journal of the Royal Meteorological Society* 81 (348): 158–171.

Jordan, T. H. 2006. Earthquake Predictability, Brick by Brick. *Seismological Research Letters* 77 (1): 3–6.

Kremer, E. 1990. On the Probable Maximum Loss. *Blätter der DGVFM* 19 (3): 201–205.

Kremer, E. 1994. More on the Probable Maximum Loss. *Blätter der DGVFM* 21 (3): 319–326.

Kuhn, M. 2021. *caret: Classification and Regression Training.* R package version 6.0-90. https://CRAN.R-project.org/package=caret.

Lamontagne, M., S. Halchuk, J. F. Cassidy, and G. C. Rogers. 2018. Significant Canadian Earthquakes 1600-2017. *Geological Survey of Canada, Open File 8285,* https://doi.org/10.4095/311183.

Lv, W., T. Mao, and T. Hu. 2012. Properties of second-order regular variation and expansions for risk concentration. *Probability in the Engineering and Informational Sciences* 26 (4): 535.

Massada, A. B., V. C. Radeloff, S. I. Stewart, and T. J. Hawbaker. 2009. Wildfire Risk in the Wildland-Urban Interface: A Simulation Study in Northwestern Wisconsin. *Forest Ecology and Management* 258 (9): 1990–1999.

McCullagh, P. 1984. Generalized Linear Models. *European Journal of Operational Research* 16 (3): 285–292.

McNeil, A. J., R. Frey, and P. Embrechts. 2015. *Quantitative Risk Management: Concepts, Techniques and Tools. Revised Edition.* Princeton University Press.

Mitchell-Wallace, K., M. Jones, J. Hillier, and M. Foote. 2017. *Natural Catastrophe Risk Management and Modelling: A Practitioner's Guide.* John Wiley & Sons.

Moller, J., and R. P. Waagepetersen. 2003. *Statistical Inference and Simulation for Spatial Point Processes.* CRC Press.

NALCMS (North American Land Change Monitoring System). 2020. *2015 Land Cover of North America at 30 Meters.* Commission for Environmental Cooperation. http://www.cec.org/nalcms.

NOAA (National Oceanic and Atmospheric Administration). 2016. "World Digital Elevation Model ETOPO5." Accessed March 15, 2021. https://www.eea.europa.eu/data-and-maps/data/world-digital-elevation-model-etopo5.

NRC (Natural Resources Canada). 2019. "Frequently Asked Questions about Earthquakes (FAQ), How often do earthquakes occur in Canada?" Accessed October 5, 2021. https://www.earthquakescanada.ca/info-gen/faq-en.php.

Ogata, Y. 1998. Space-Time Point-Process Models for Earthquake Occurrences. *Annals of the Institute of Statistical Mathematics* 50 (2): 379–402.

Ohlsson, E., and B. Johansson. 2010. *Non-life Insurance Pricing with Generalized Linear Models.* Vol. 174. Springer.

Onur, T., C. E. Ventura, and W. D. L. Finn. 2005. Regional Seismic Risk in British Columbia - Damage and Loss Distribution in Victoria and Vancouver. *Canadian Journal of Civil Engineering* 32 (2): 361–371.

OSFI (Office of the Superintendent of Financial Institutions). 2019. *Minimum Capital Test For Federally Regulated Property and Casualty Insurance Companies.* https://www.osfi-bsif.gc.ca/Eng/fi-if/rg-ro/gdn-ort/gl-ld/Pages/mct2019.aspx.

Pickands III, J. 1975. Statistical Inference Using Extreme Order Statistics. *The Annals of Statistics* 3 (1): 119–131.

Pimentel, R. S., M. Niewiadomska-Bugaj, and J.-C. Wang. 2015. Association of Zero-inflated Continuous Variables. *Statistics & Probability Letters* 96:61–67.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rhoades, D. A., D. Schorlemmer, M. C. Gerstenberger, A. Christophersen, J. D. Zechar, and M. Imoto. 2011. Efficient Testing of Earthquake Forecasting Models. *Acta Geophysica* 59 (4): 728–747.

Rowlingson, B., and P. Diggle. 2022. *splancs: Spatial and Space-Time Point Pattern Analysis.* R package version 2.01-43. https://CRAN.R-project.org/package=splancs.

Schelldorfer, J., and M. V. Wüthrich. 2019. Nesting Classical Actuarial Models into Neural Networks. *Available at SSRN 3320525.*

Schmeidler, D. 1986. Integral Representation Without Additivity. *Proceedings of the American mathematical society* 97 (2): 255–261.

Schmeidler, D. 1989. Subjective Probability and Expected Utility without Additivity. *Econometrica: Journal of the Econometric Society,* 571–587.

Schorlemmer, D., M. Gerstenberger, S. Wiemer, D. Jackson, and D. Rhoades. 2007. Earthquake Likelihood Model Testing. *Seismological Research Letters* 78 (1): 17–29.

Schorlemmer, D., and M. Gerstenberger. 2007. RELM Testing Center. *Seismological Research Letters* 78 (1): 30–36.

Shi, P., and K. Shi. 2017. Territorial Risk Classification Using Spatially Dependent Frequency-Severity Models. *ASTIN Bulletin: The Journal of the IAA* 47 (2): 437–465.

Short, K. C. 2017. *Spatial Wildfire Occurrence Data for the United States, 1992-2015 [FPA_FOD_20170508].* 4th ed. Fort Collins, CO: Forest Service Research Data Archive. Accessed December 15, 2020. https://doi.org/10.2737/RDS-2013-0009.4.

Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis.* Vol. 26. CRC Press.

Solvency II. *Commission Delegated Regulation (EU) 2015/ 35 of 10 October 2014 Supplementing Directive 2009/138/EC of the European Parliament and of the Council on the Taking-up and Pursuit of the Business of Insurance and Reinsurance (Solvency II).* https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32015R0035.

Statistics Canada, Government of Canada. 2016. "Census Profile, 2016 Census." Accessed August 15, 2020.

Statistics Canada, Government of Canada. 2017. "Population Ecumene Census Division Cartographic Boundary File, Reference Guide, 2016 Census." Accessed August 15, 2020. https://www150.statcan.gc.ca/n1/pub/92-159-g/92-159-g2016001-eng.htm.

Statistics Canada, Government of Canada. 2020a. "Canadian Housing Statistics Program (CHSP)." Accessed August 15, 2020. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5257.

Statistics Canada, Government of Canada. 2020b. "Table 18-10-0135-02 Building Construction Price Indexes, Percentage Change, Quarterly." Accessed August 15, 2020. https://www.statcan.gc.ca/eng/survey/business/2317.

Statistics Canada, Government of Canada. 2020c. "Table 34-10-0066-01 Building Permits, by Type of Structure and Type of Work." Accessed August 15, 2020. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3410006601.

Stocks, B., J. Mason, J. Todd, E. Bosch, B. Wotton, B. Amiro, M. Flannigan, K. Hirsch, K. Logan, D. Martell, et al. 2002. Large Forest Fires in Canada, 1959–1997. *Journal of Geophysical Research: Atmospheres* 107 (D1): FFR–5.

Swiss Re. 2017. "Earthquake Risk in Eastern Canada: Mind the Shakes." Accessed August 19, 2021. https://www.swissre.com/institute/research/topics-and-risk-dialogues/climate-and-natural-catastrophe-risk/earthquake-risk-in-eastern-canada-mind-the-shakes.html.

Thibert, K. M. 2008. "A Methodology for Assessing the Seismic Risk of Buildings." PhD diss., University of British Columbia.

Troop, D., F. Godin, and J. Y. Yu. 2021. "Bias-corrected peaks-over-threshold estimation of the CVaR." In *Uncertainty in Artificial Intelligence,* 1809–1818. PMLR.

Turner, R. 2021. *deldir: Delaunay Triangulation and Dirichlet (Voronoi) Tessellation.* R package version 1.0-6. https://CRAN.R-project.org/package=deldir.

U.S. Geological Survey. 2018. "East vs West Coast Earthquakes: U.S. Geological Survey." Accessed September 30, 2021. https://www.usgs.gov/news/featured-story/east-vs-west-coast-earthquakes.

U.S. Geological Survey. 2021. "Cool Earthquake Facts." Accessed September 27, 2021. https://www.usgs.gov/natural-hazards/earthquake-hazards/science/cool-earthquake-facts.

U.S. Geological Survey. 2022a. "Can You Predict Earthquakes?" Accessed February 15, 2022. https://www.usgs.gov/faqs/can-you-predict-earthquakes.

U.S. Geological Survey. 2022b. "Where do earthquakes occur?" Accessed April 25, 2022. https://www.usgs.gov/faqs/where-do-earthquakes-occur.

Ulmi, M., C. L. Wagner, M. Wojtarowicz, J. L. Bancroft, N. L. Hastings, W. Chow, J. R. Rivard, et al. 2014. HAZUS-MH 2.1 Canada User and Technical Manual: Earthquake Module. *Geological Survey of Canada, Open File 7474,* https://doi.org/10.4095/293800.

Van Wagtendonk, J. W. 1996. "Use of a Deterministic Fire Growth Model to Test Fuel Treatments." In *Sierra Nevada Ecosystem Project, Final Report to Congress,* 2:1155–1166.

Vandewalle, B., and J. Beirlant. 2006. On univariate extreme value statistics and the estimation of reinsurance premiums. *Insurance: Mathematics and Economics* 38 (3): 441–459.

Von Mises, R. 1954. *La distribution de la plus grande de n valeurs.* II:271–294. American Mathematical Society, Providence, RI.

Wald, D. J., V. Quitoriano, T. H. Heaton, and H. Kanamori. 1999. Relationships between Peak Ground Acceleration, Peak Ground Velocity, and Modified Mercalli Intensity in California. *Earthquake Spectra* 15 (3): 557–564.

Wallemacq, P., and R. House. 2017. Economic Losses, Poverty and Disasters 1998-2017. *United Nations Office for Disaster Risk Reduction.*

Wang, S. 1995. Insurance pricing and increased limits ratemaking by proportional hazards transforms. *Insurance: Mathematics and Economics* 17 (1): 43–54.

Wang, S. S. 2000. A Class of Distortion Operators for Pricing Financial and Insurance Risks. *Journal of Risk and Insurance,* 15–36.

Westerling, A. L., H. G. Hidalgo, D. R. Cayan, and T. W. Swetnam. 2006. Warming and Earlier Spring Increase Eestern US Forest Wildfire Activity. *Science* 313 (5789): 940–943.

Wilkinson, M. E. 1982. "Estimating Probable Maximum Loss with Order Statistics." In *Casualty Actuarial Society Forum,* 195–209.

Wong, K., and F. P. Schoenberg. 2009. On Mainshock Focal Mechanisms and the Spatial Distribution of Aftershocks. *Bulletin of the Seismological Society of America* 99 (6): 3402–3412.

Wood, H. O., and F. Neumann. 1931. Modified Mercalli Intensity Scale of 1931. *Bulletin of the Seismological Society of America* 21 (4): 277–283.

Wüthrich, M. V. 2018. Machine Learning in Individual Claims Reserving. *Scandinavian Actuarial Journal* 2018 (6): 465–480.

Wüthrich, M. V., and C. Buser. 2020. Data Analytics for Non-life Insurance Pricing. *Swiss Finance Institute Research Paper,* nos. 16-68.

Yaghmaei, N. 2020. *Human Cost of Disasters: An Overview of the Last 20 Years, 2000-2019.* United Nations Office for Disaster Risk Reduction.

Yang, F. 2015. First- and second-order asymptotics for the tail distortion risk measure of extreme risks. *Communications in Statistics. Theory and Methods* 44 (3): 520–532. ISSN: 0361-0926. https://doi.org/10.1080/03610926.2012.751116. https://doi.org/10.1080/03610926.2012.751116.

Zechar, J. D., M. C. Gerstenberger, and D. A. Rhoades. 2010. Likelihood-based Tests for Evaluating Space-rate-magnitude Earthquake Forecasts. *Bulletin of the Seismological Society of America* 100 (3): 1184–1195.

Zhao, Y., T. Mao, and F. Yang. 2021. Estimation of the Haezendonck-Goovaerts risk measure for extreme risks. *Scandinavian Actuarial Journal,* 1–24.

Zhu, L., and H. Li. 2012. Tail distortion risk and its asymptotic analysis. *Insurance: Mathematics & Economics* 51 (1): 115–121. ISSN: 0167-6687. https://doi.org/10.1016/j.insmatheco.2012.03.010. https://doi.org/10.1016/j.insmatheco.2012.03.010.

Zhuang, J. 2006. Second-order Residual Analysis of Spatiotemporal Point Processes and Applications in Model Evaluation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68 (4): 635–653.

Zhuang, J., Y. Ogata, and D. Vere-Jones. 2002. Stochastic Declustering of Space-time Earthquake Occurrences. *Journal of the American Statistical Association* 97 (458): 369–380.