



UNIL | Université de Lausanne

Unicentre

CH-1015 Lausanne

<http://serval.unil.ch>

Year : 2019

Two specific problems in Data Science: Demand forecasting using weather data and Non-linear causality inference

Babongo Bosombo Flora

Babongo Bosombo Flora, 2019, Two specific problems in Data Science: Demand forecasting using weather data and Non-linear causality inference

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive <http://serval.unil.ch>

Document URN : urn:nbn:ch:serval-BIB_35E2A0F511042

Droits d'auteur

L'Université de Lausanne attire expressément l'attention des utilisateurs sur le fait que tous les documents publiés dans l'Archive SERVAL sont protégés par le droit d'auteur, conformément à la loi fédérale sur le droit d'auteur et les droits voisins (LDA). A ce titre, il est indispensable d'obtenir le consentement préalable de l'auteur et/ou de l'éditeur avant toute utilisation d'une oeuvre ou d'une partie d'une oeuvre ne relevant pas d'une utilisation à des fins personnelles au sens de la LDA (art. 19, al. 1 lettre a). A défaut, tout contrevenant s'expose aux sanctions prévues par cette loi. Nous déclinons toute responsabilité en la matière.

Copyright

The University of Lausanne expressly draws the attention of users to the fact that all documents published in the SERVAL Archive are protected by copyright in accordance with federal law on copyright and similar rights (LDA). Accordingly it is indispensable to obtain prior consent from the author and/or publisher before any use of a work or part of a work for purposes other than personal use within the meaning of LDA (art. 19, para. 1 letter a). Failure to do so will expose offenders to the sanctions laid down by this law. We accept no liability in this respect.



UNIL | Université de Lausanne

FACULTÉ DES HAUTES ÉTUDES COMMERCIALES
DÉPARTEMENT DES SYSTÈMES D'INFORMATION

**Two specific problems in Data Science:
Demand forecasting using weather data
and
Non-linear causality inference**

THÈSE DE DOCTORAT

présentée à la

Faculté des Hautes Études Commerciales
de l'Université de Lausanne

pour l'obtention du grade de
Docteur ès Sciences en systèmes d'information

par

Flora BABONGO BOSOMBO

Directeur de thèse
Prof. Ari-Pekka Hameri

Co-directrice de thèse
Prof. Valérie Chavez-Demoulin

Jury

Prof. Felicitas Morhart, Présidente
Prof. Olivier Gallay, expert interne
Prof. Ralf Seifert, expert externe

LAUSANNE
2019

IMPRIMATUR

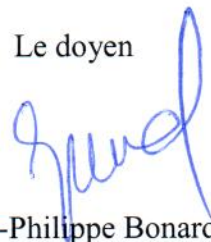
Sans se prononcer sur les opinions de l'autrice, la Faculté des Hautes Etudes Commerciales de l'Université de Lausanne autorise l'impression de la thèse de Madame Flora BABONGO BOSOMBO, titulaire d'un bachelor en Economie Politique de l'Université de Lausanne, et d'un master en Statistique de l'Université de Neuchâtel, en vue de l'obtention du grade de docteur ès Sciences en systèmes d'information.

La thèse est intitulée :

**TWO SPECIFIC PROBLEMS IN DATA SCIENCE:
DEMAND FORECASTING USING WEATHER DATA
AND
NON-LINEAR CAUSALITY INFERENCE**

Lausanne, le 09 juillet 2019

Le doyen



Jean-Philippe Bonardi



UNIL | Université de Lausanne

THESIS COMMITTEE:

Ari-Pekka HAMERI

Thesis supervisor, Professor of Operations management, Faculty of Business and Economics (HEC), University of Lausanne

Valérie CHAVEZ-DEMOULIN

Thesis co-supervisor, Professor of Statistics, Faculty of Business and Economics (HEC), University of Lausanne

Felicitas MORHART

President of the jury, Professor of Marketing, Faculty of Business and Economics (HEC), University of Lausanne

Olivier GALLAY

Internal expert, Professor of Statistics, Faculty of Business and Economics (HEC), University of Lausanne

Ralf SEIFERT

External expert, Professor of Mathematical models in supply chain management, Federal Polytechnic School of Lausanne EPFL

University of Lausanne
Faculty of Business and Economics


PhD in Information Systems

I hereby certify that I have examined the doctoral thesis of

Flora BABONGO BOSOMBO

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or committee members
made during the doctoral colloquium
have been addressed to my entire satisfaction.

Signature:  Date: 05.07.2019

Prof. Ari-Pekka HAMERI
Thesis supervisor

University of Lausanne
Faculty of Business and Economics

PhD in Information Systems

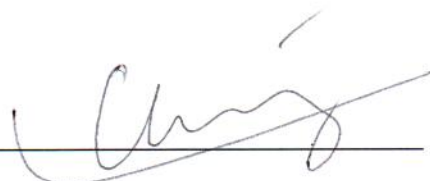
I hereby certify that I have examined the doctoral thesis of

Flora BABONGO BOSOMBO

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or committee members
made during the doctoral colloquium
have been addressed to my entire satisfaction.

Signature: _____



Date: _____

05.07.2019

Prof. Valérie CHAVEZ-DEMOULIN
Thesis co-supervisor

University of Lausanne
Faculty of Business and Economics

PhD in Information Systems

I hereby certify that I have examined the doctoral thesis of

Flora BABONGO BOSOMBO

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or committee members
made during the doctoral colloquium
have been addressed to my entire satisfaction.

Signature: _____



Date: _____

05.07.2019

Prof. Olivier GALLAY
Internal member of the doctoral committee

University of Lausanne
Faculty of Business and Economics

PhD in Information Systems

I hereby certify that I have examined the doctoral thesis of

Flora BABONGO BOSOMBO

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or committee members
made during the doctoral colloquium
have been addressed to my entire satisfaction.

Signature:  Date: 05.07.2019

Prof. Ralf SEIFERT
External member of the doctoral committee

Dedication

I dedicate this thesis to all my family , especially

- To my elder sister **Judith BABONGO** and her husband **Pierrot MUNDEMBA**
- To my godparents **Christoph SOLAND** and **Marie-Françoise PIOT**

their unconditional love and support have been an integral part of the completion of this thesis, and I am eternally grateful for everything they have done for me.

I also dedicate this thesis to the loving memory of my daughter, **Emma**.

Acknowledgement

First and foremost, I wish to thank my advisors, Ari-Pekka Hameri and Valérie Chavez-Demoulin, for their mentoring and trust. They offered me their helpful guidance and advice which helped me to grow further as an academic researcher and as a person.

A special thanks to Tapio Niemi, co-author in 3 of the papers developed in this thesis, for his comments and discussions that significantly contributed to improve my research.

I also would like to acknowledge my committee members for their time and expertise.

I am grateful to the University of Lausanne, which enabled me to work under perfect conditions. Additionally I want to thank the Swiss National Science Foundation for their financial support.

I thank all my friends and colleagues, members of the Department of Operations and of the Department of Information Systems.

Finally I would like to thank my family, who has always been there to listen, support and encourage me. They were emotionally an infinite support.

Abstract

In this thesis, I investigate two specific subjects in data science, namely demand forecasting and causality inference, dividing this thesis in two main parts.

The first part aims at improving demand forecasting accuracy that impacts supply chain performance. It consists of three articles aiming at studying how to enhance demand forecasting accuracy using pertinent data (e.g. operational transaction data, weather data, socio-economic data, etc.). Each article explores a new statistical approach on the supply chain optimization through demand forecasting accuracy.

- In the first article we analyze transactional longitudinal data of several business units, matched with daily location-based weather conditions. We also study ways in which weather fluctuations affect supply chain performance through the delivery delay in days. Understanding this relationship is valuable both for improving sales forecast accuracy and for improving operational performance.
- The second article aims at explaining how weather conditions and fluctuations affect the accuracy of demand forecasting for seasonal products. We found that weather conditions have a significant impact on demand forecasting accuracy with reductions in percentage errors up to 45%. These results can be used to justify and motivate the integration of the impact of variability in weather in the decision making process in order to better anticipate demand volumes and reduce costs due to excess inventory or stock shortages.
- The goal of the third article is to improve demand forecasting accuracy by using the concept of spatial dependence and interpolation, and incorporating the effects of socio-economic aspects and weather conditions in the spatial dependence structure. The accuracy of demand forecasting is improved, the reduction of the forecasting error is up to 48%.

The goal of the second part is to infer the causal relationship in the case of non-linearity and heteroscedasticity.

- In the fourth article, a two-steps method is proposed to infer the intrinsic causal mechanism between two variables dealing with heteroscedasticity. We provide a bivariate multiplicative noise model that we extend to the multiplicative case. The two-steps Causal Heteroscedastic Model consists of applying a causal additive model on the BAMLSS (bayesian additive model for location, scale and shape) fitted values of the estimated parameters. The simulation study provides an accuracy of 0.97 on average.

In this thesis, I have explored and analyzed two specific subjects in data science, which are demand forecasting and non-linear causality inference. This thesis has provided several studies improving demand forecasting accuracy by reducing the forecasting error in several contexts dealing with seasonality, through the integration of external data such as weather or socio-economic data, using complex statistical models. The causal method provided in this thesis allows the inference of inherent causal mechanism.

Résumé

Dans cette thèse j'investigue deux sujets particuliers de la science des données, à savoir la prévision de la demande et l'inférence de la causalité, divisant cette thèse en deux parties.

Le but de la première partie est d'améliorer la précision de la prévision de la demande car elle impacte la performance de la chaîne logistique. Cette partie comprend trois articles dans lesquels nous étudions comment améliorer la précision des prévisions de la demande grâce à l'incorporation des données pertinentes dans le modèle d'analyse. Chacun des trois articles explore une nouvelle approche statistique.

- Dans le premier article, nous analysons les données transactionnelles des opérations de plusieurs unités commerciales, jumelées avec les données sur les conditions météorologiques journalières. Nous analysons aussi comment les fluctuations de la météo affectent la performance de la chaîne logistique. La compréhension de ces relations est importante et utile pour l'amélioration de la précision des prévisions de la demande.
- Le but du deuxième article est d'analyser et d'expliquer comment les conditions météorologiques ainsi que ses fluctuations impactent la précision des prévisions de la demande saisonnière. Les résultats montrent que le temps qu'il fait a un impact significatif sur cette précision, réduisant le pourcentage d'erreur de 45%. Ces résultats peuvent être utilisés pour justifier et motiver l'intégration de l'impact de la météo dans le processus décisionnel.
- Le troisième article utilise la dépendance spatiale pour améliorer la précision des prévisions de la demande, ainsi que l'incorporation des effets des facteurs socio-économiques et des conditions météorologiques dans la structure de cette dépendance spatiale. Les résultats révèlent une amélioration de la précision et une réduction de l'erreur de prédiction allant jusqu'à 48%.

La deuxième partie de cette thèse explore l'inférence de la causalité dans le cas de la non-linéarité et de l'hétéroscédasticité.

- Dans le quatrième article, nous proposons une méthode à deux étapes pour inférer le mécanisme causal intrinsèque entre deux variables en présence d'hétéroscédasticité. Nous proposons un modèle bivarié et multiplicatif par rapport au terme d'erreur que nous étendons au cas multivarié ensuite. Le modèle à deux étapes appelé Causal Heteroscedastic Model (CHM) consiste à appliquer un CAM (causal additive model) aux valeurs ajustées des paramètres estimés par un modèle BAMLSS (bayesian additive model for location, scale and shape). Les simulations effectuées montrent que le CHM trouve la bonne causalité dans 97% des cas en moyenne.

Dans cette thèse, j'ai exploré et analysé deux sujets spécifiques de la science des données, qui sont la prévision de la demande et l'inférence de la causalité non-linéaire. Cette thèse comprend plusieurs études améliorant la précision des prévisions de la demande, dans différents contextes comme la saisonnalité, en réduisant l'erreur de prédiction grâce aux données pertinentes et aux outils statistiques complexes. Quant au modèle à deux étapes proposé, il permet l'inférence du mécanisme inhérent de la causalité.

Contents

Dedication	i
Acknowledgment	iii
Abstract	v
Résumé	vii
1 Introduction	1
Part I: Demand forecasting using weather	6
2 Weather and supply chain performance in sport goods distribution	9
2.1 Introduction	11
2.2 Literature review on weather affecting operations and supply chain management	13
2.3 Research hypothesis; data and methodology	15
2.3.1 Building up detailed research hypotheses	16
2.3.2 The case company and data	17
2.3.3 Explaining weather effect for demand	20
2.3.4 Modelling delays by using weather information	23
2.4 Further results and hypothesis validity	24
2.5 Discussion and future research	29
2.6 References	31
2.7 Appendix	34
3 Using weather data to improve demand forecasting for seasonal products	37
3.1 Introduction	40
3.2 Literature review	42
3.3 Motivation; research questions; data and methodology	46
3.3.1 Motivation and research questions	46
3.3.2 Data	46
3.3.3 Methodology	48
3.4 Results	52
3.4.1 Model results	52
3.4.2 Prediction results	55
3.5 Discussion	56
3.6 Conclusions and future research	57

3.7	References	58
3.8	Notes	62
4	Forecasting (un-)seasonal demand using geostatistics, socio-economic and weather data	63
4.1	Introduction	66
4.2	Literature review	68
4.2.1	Socio-economic environment and weather conditions in demand forecasting	68
4.2.2	Geostatistics applications in various fields	68
4.2.3	Geostatistics applied to demand forecasting	69
4.3	Research questions; data and methodology	70
4.3.1	Research questions	70
4.3.2	Data	70
4.3.3	Methodology	74
4.4	Results	76
4.4.1	Semivariograms	76
4.4.2	Model fitting results	77
4.4.3	Prediction results	80
4.5	Conclusions	83
4.6	References	84
4.7	Notes	86
	Part II: Non-linear causal inference	87
5	Causal discovery for heteroscedastic financial series	89
5.1	Introduction	89
5.2	Causal discovery for heteroscedastic model	91
5.2.1	First step: BAMLSS	91
5.2.2	Second step: Bivariate CAM	93
5.3	Simulation study	94
5.4	Stock market indices	101
5.4.1	Pairwise exploration	101
5.4.2	Extension to multivariate case	101
5.5	Conclusion	104
5.6	Appendix	105
	Bibliography	109

Chapter 1

Introduction

In this thesis, I investigate two specific issues in data science, namely demand forecasting and causality inference. Let's start by defining data science as an interdisciplinary field combining statistics and computer science, aiming to understand and analyze actual phenomena through large databases. According to [Van der Aalst \[2016\]](#), data science includes data extraction, preparation, exploration, transformation, data storage and retrieval, computing infrastructures, various types of mining and learning, presentation of explanations and predictions. Data science aims at providing meaningful information based on large amounts of complex data for decision-making purposes. Nowadays, data science is omnipresent in everyone's everyday life. Applications of data science are numerous. For example, receiving personalized advertisements related to our past online researches, the fact that YouTube shows all our favorite videos on our home screen, or healthcare parameters control through connected watches etc.

Moreover, in recent years, data science has become an essential component of many industries and fields.

- In biomedicine the application of data science on biomedical information such as human genome, provides an opportunity for personalized medicine programs in order to improve patient care through modern sequencing technology allowing high-resolution genetic sequencing at tremendous scale [[Costa, 2014](#)].
- In social science, professionals have now access to terabytes of data describing almost instantly human behavior and interactions between individuals, this allows them to study and try to understand contours of society through data science methodologies. For instance, [Moussaïd et al. \[2011\]](#) study pedestrian flows and crowd disasters. They suggest that, guided by visual information such as the distance of obstructions in individual lines of sight, pedestrians adapt their walking speeds and directions. Their model aims to predict individual trajectories and collective patterns of motion.
- To enter a market with specific characteristics, marketers have to segment the potential customers and understand their needs. For this purpose, data science is used to develop predictive and descriptive methods such as clustering techniques. Data science is also applied for more general

customer relationship management, for example customer behavior analyses in order to maximize expected customer value [Provost and Fawcett, 2013].

- Financial institutions were among the early users of data science. Data science is applied in numerous domains such as portfolio management, investment risk analysis, prediction of bankruptcy, etc. The use of data science helps the banks better understand customer needs and anticipate their response to new products and services [Dick, 2008].

Numerous other fields are impacted by data science. Our interest in data science is motivated by a desire to improve demand forecasting accuracy impacting supply chain performance on one hand, and to infer the causal relationship in the case of non-linearity and heteroscedasticity on the other hand, dividing this thesis in two parts.

Part 1: Demand forecasting

Supply chain management (SCM) plays an essential role in corporate efficiency. It consists of designing, planning, monitoring and optimizing supply chain activities, from supplying raw materials to delivering final products, in order to create net value and synchronize supply with demand. One way of achieving the goal of SCM is to minimize total costs with respect to frictions of different chain partners, for example considering the inventory level, the sale department will tend to opt for higher inventory levels in order to fulfill demands whereas the warehouse division will prefer lower inventories so as to reduce storage costs [Ayers, 2006; Sadeghi et al., 2016]. For the purpose of matching supply with demand, demand forecasting is a fundamental component of supply chain process. Historical operational and sales data are utilized to estimate the expected forecast of customer demand that are used for almost all supply chain related decisions such as:

- Optimization of inventory levels: the most accurate demand forecast allows an optimized management through right decisions concerning the whole process going from desired raw material to finished goods and their inventory level [Singh and Kumar, 2011].
- Customer service/satisfaction level: proper forecast of customer demand helps to adapt the offering to a wide variety of customers. Indeed, understanding the customer's situation and need contributes to superior demand chain efficiency and high customer satisfaction [Heikkilä, 2002].

Demand forecasting has been widely studied in both qualitative and quantitative reasoning approaches. Hofmann and Rutschmann [2018] showed that demand forecasting is a complicated task that could benefit from additional relevant data and processes and they examine how big data analytics improve the accuracy of demand forecasts. They found that the integration of different data sources in demand forecasting process is feasible but requires data scientists and appropriate technology investments. Hence the first part of this thesis consists studying how to enhance demand forecasting accuracy using pertinent data (e. g. operational transaction data, weather data, socio-economic data, etc.). Each article explores a new statistical approach on the supply chain optimization through demand forecasting accuracy.

Article 1:

Uncertainty has been proved to negatively affect supply chain performance [Dahistrom et al., 1996; Morris and Carter, 2005]. The starting point of the improvement of supply chain performance is to understand the customer's demands in order to optimize the asset utilization, to eliminate the excess inventories costs and to reduce lead time. For this purpose, several studies focus on the impacts of extreme weather [Tierney, 1997; Blackhurst et al., 2011]. The effects of weather on productivity has been more investigated than its effects on supply chain performance, especially in agricultural and construction industries [Thomas et al., 1999]. Since weather plays a role in the operational activities, the first article consists of studying how everyday weather fluctuations impact supply chain performance in different business of sport goods.

We analyze several business units with different operational strategies through real transactional business longitudinal data matched with daily location-based weather conditions (temperature, quantity of snow/rain, length of sunshine per day). We found that when the temperature increases the mean order volume significantly decreases for small customers in resorts ordering seasonal products in winter. In other words, weather at customer locations has a significant effect on order volumes and this effect differs according to the type of product, the location and the size of the customer; and non-urban locations or resorts seem more vulnerable than urban areas to weather variability. We also study ways in which weather fluctuations affect supply chain performance, that is, the delivery delay in days. We found that when the temperature increases, the delay in days also significantly decreases.

We were also interested in analyzing how weather fluctuations affect the dependence between order volume and delays. We found that order volume and delay are more dependent during winter than during summer.

The results of this article can be used to estimate and explain the weather effect in supply chain performance. Understanding this relationship is valuable both for improving sales forecast accuracy and for improving operational performance.

Article 2:

Demand forecasts play a crucial role for supply chain management especially in case of seasonal products because of the conflict opposing retailers to manufacturers concerning the order time. Indeed, due to the demand uncertainty of seasonal products, retailers tend to place their orders as late as possible in order to gather more information and reduce demand forecasting error, whereas manufacturers having limited productions capacity wish to have orders as soon as possible [Chen and Xu, 2001]. According to Chen and Yano [2010], weather is an important determinant of demand for the seasonal products. In the second article we aim to explain how weather conditions and fluctuations affect the accuracy of demand forecasting for seasonal products, namely winter sport goods which are ordered and manufactured over 8 months before to be sold to customers, meaning a long lag between ordering and delivery. We analyze real transaction business data of alpine ski products of different brands matched with daily location based weather conditions.

We found that weather conditions have a significant impact on demand forecasting accuracy. The incremental improvement gained is the reduction in percentage errors up to 45%. The contribution of this article is the operationalization of a ‘great winter’ and the demonstration of the fact that weather in one winter affects sales in the next winter. These results can be used to justify and motivate the integration of the impact of variability in weather in the decision making process in order to better anticipate demand volumes and reduce costs due to excess inventory or stock shortages.

Article 3:

Seasonal products are common in many industries and can involve a large number of factors such as the influence of seasonal weather changes or socio-economic features. Since the weather characteristics such as temperature or precipitation, and socio-economics features are spatially dependent [Ashraf et al., 1997; Anselin, 1999], we assume that close customers are more likely to similar demand according to weather and socio-economic features and customers far apart from each other are more likely to have less similar demand.

Numerous studies of seasonal products are based on time series statistical techniques [Adhikari and Agrawal, 2012; Gan et al., 2014] and less on geostatistics. Geostatistics have been applied to model the spatial dependence in various fields such as mining industry, soil science, agriculture etc.

Assuming that customers in a neighborhood may imitate each other leading to spatial dependence, we aim in the third article to improve demand forecasting accuracy by using the concept of spatial dependence and interpolation, and incorporating the effects of socio-economic aspects and weather conditions in the spatial dependence structure. We focus on studying the demand fluctuation of seasonal (winter sport goods) and unseasonal leisure goods (indoor sports and golf equipment). We analyse real demand data to find first how it varies geographically according to socio-economic aspects and weather conditions; and second how the additional information, i.e external to the supply chain, affects demand forecasting accuracy.

As main results we show that weather conditions impact the spatial correlation of the demand of seasonal products, but they do not have a significant impact for unseasonal products. We found that socio-economic features impact spatial correlation of both seasonal and unseasonal demand. The accuracy of demand forecasting is improved by the incorporation of weather conditions and socio-economic features in the forecasting process, the reduction of the forecasting error is up to 48%.

These results can be used in the decision making process, for example for planning future demand in order optimize inventories and orders, or for deciding the location of a new retail shop.

Part 2: Causality inference

It is well known to anyone who has basic notions of statistics that "correlation does not mean causation". The most famous example is the link between country’s chocolate consumption and Nobel Prize victories [Messerli, 2012]. This article provides a graph showing a strong correlation between chocolate consumption per capita in a country and the number of Nobel laureates per

capita as well. Since this correlation does not imply causation, we have three possibilities: Either chocolate influences Nobel price or the opposite, or both chocolate consumption and Nobel price are influenced by a common underlying mechanism such as the country's economic features and the investment capacity in research. Therefore, the best way to go from correlation to causality is, identifying causal relationships from controlled randomized experiments [Rubin, 1974], but these experiments are in many cases too costly or unethical and even infeasible.

Inferring causality from observational data is one of the fundamental subjects in empirical science. The alternative developed tools to controlled randomized experiments, are based on inferring causal relationships from observational data using conditional independence [Rubin, 1974; Pearl, 2009; Spirtes and Zhang, 2016]. In the bivariate case when observing only two variables (X and Y), causality inference consists of identifying the direct causation " $X \rightarrow Y$ " or " $X \leftarrow Y$ " with the assumption that there is no latent confounding variable causing both X and Y . In the additive context, this problem has been studied by imposing certain model specifications or restrictions. For linear causal models ($Y = bX + \varepsilon$), if at most one of X and ε is gaussian, the causal direction is identifiable, due to the independent component analysis (ICA) theory [Hyvärinen et al., 2004]. The linear non-Gaussian causal model, known as LinGAM [Shimizu et al., 2006] also relies on ICA with the additional assumption that disturbance variables have non-gaussian distributions of non-zero variances. Even though linear causal models with additive noise are often used because they are well understood and there are well-known methods, nevertheless in reality many causal relationships are more or less nonlinear. Nonlinearities in the data-generating process provide more information on the underlying causal system since these models allow more aspects of the true data generating mechanisms to be identified ($Y = f(X) + \varepsilon$) [Hoyer et al., 2009]. The post-nonlinear causal model ($Y = g(f(X) + \varepsilon)$) provided by [Zhang and Hyvärinen, 2009] aims to distinguish the cause from effect by analyzing the nonlinear effect of the cause, the inner noise effect, and the measurement distortion effect in the observed variables. According to the literature review, most of the papers analyze additive models with either linearity, nonlinearity or gaussian noise, the case of a nonlinear and non-gaussian causal multiplicative noise model ($Y = f(X) + g(X)\varepsilon$) has been less explored.

In finance, causality is mostly studied conditioned on time. For example, the linear and nonlinear intertemporal cross correlation [Atchison et al., 1987] aims to infer causality according to time. This method relies on the fact that asset prices change in a time-lag manner and not simultaneously. In other words, price-adjustment delay factors along with nonsynchronous trading cause the autocorrelations present in daily asset returns.

The most explored is the widely used Granger causality [Granger and Morgenstern, 1963]. Under Granger causality, the cause happens prior to its effect. It aims to determine whether one time series is useful in forecasting another. A time series X is said to Granger-cause Y if it can be tested that lagged values of X provide statistically significant information about future values of Y . Granger [1981] provides a cointegrated form causality based on the fact that, in finance, assets can move in an integrated manner, meaning that they evolve dynamically together and this joint evolution can be analyzed as a linear or

nonlinear integrated dependencies function. More precisely two time series are considered as cointegrated when their combination is stationary. For the linear case if asset X is negatively cointegrated with asset Y , this means that if the price of asset X increases at time $t - 1$, then the price of asset Y shall decrease at time t . Besides the up cited methods, numerous time series models aim to infer causal dependencies nevertheless they are all conditioned on time, hence one cannot observe assets and infer the causality simultaneously.

Article 4:

Inferring causality between financial assets is a common and fundamental subject in finance. We have seen that most of the existing methods are conditioned on time. In this paper, we aim to infer the intrinsic causal mechanism between two financial heteroscedastic time series. Unlike the Granger causality which infers only at the mean level, we investigate causal relations not only in mean but from the perspective of location, scale and shape parameters of the underlying distribution. We propose a new two-steps method Causal Heteroscedastic Model (CHM) that is not conditioned on time and can handle any response distribution since it infers the inherent causality through all the parameters of the underlying distribution. We focus on the bivariate multiplicative noise model $Y = f(X) + g(X)\varepsilon$. The two-steps CHM consists of applying a causal additive model (CAM) on the BAMLSS (bayesian additive model for location, scale and shape) fitted values of the estimated parameters. We have tested our method on both simulated and real financial indices log-returns data. We found that CHM reaches the accuracy of 0.97 on average. On financial data we fitted both bivariate and the multivariate CHM, we find an intrinsic causal effect of the shares on the index they compose. The multivariate analysis provides directed acyclic graphs (DAG) revealing the causal structure between shares in normal and extreme case. This new method is a real contribution to causality research since it can deal with any response distribution and it is applicable to many other domains in future research, such as genomics etc.

Conclusion

In this thesis, I have explored and analyzed two specific subjects in data science, which are Demand forecasting and Causality inference. With proper demand forecasting, supply chain and business performance can be considerably improved, resulting in numerous benefits, such lead time reduction, storage costs reduction and more important customer satisfaction. This thesis has provided several studies improving demand forecasting accuracy by reducing the forecasting error in several contexts dealing with seasonality, through the integration of external data such as weather or socio-economic data, using complex statistical models.

The causal method provided in this thesis allows the inference of inherent causal mechanism between assets unconditioned on time. The developed Causal Heteroscedastic Model is applied to financial index data highlighting ground-truth causal evidence and opens wide the door to numerous other applications in finance or any other domain dealing with heteroscedasticity.

Part I: Demand forecasting using weather

Chapter 2

Weather and supply chain performance in sport goods distribution

Weather and supply chain performance in sport goods distribution

Patrik Appelqvist
Amer Sports Corporation, Helsinki, Finland, and
Flora Babongo, Valérie Chavez-Demoulin,
Ari-Pekka Hameri and Tapio Niemi
*Faculty of Business and Economics, University of Lausanne,
Lausanne, Switzerland*

Abstract

Purpose – The purpose of this paper is to study how variations in weather affect demand and supply chain performance in sport goods. The study includes several brands differing in supply chain structure, product variety and seasonality.

Design/methodology/approach – Longitudinal data on supply chain transactions and customer weather conditions are analysed. The underlying hypothesis is that changes in weather affect demand, which in turn impacts supply chain performance.

Findings – In general, an increase in temperature in winter and spring decreases order volumes in resorts, while for larger customers in urban locations order volumes increase. Further, an increase in volumes of non-seasonal products reduces delays in deliveries, but for seasonal products the effect is opposite. In all, weather affects demand, lower volumes do not generally improve supply chain performance, but larger volumes can make it worse. The analysis shows that the dependence structure between demand and delay is time varying and is affected by weather conditions.

Research limitations/implications – The study concerns one country and leisure goods, which can limit its generalizability.

Practical/implications – Well-managed supply chains should prepare for demand fluctuations caused by weather changes. Weekly weather forecasts could be used when planning operations for product families to improve supply chain performance.

Originality/value – The study focuses on supply chain vulnerability in normal weather conditions while most of the existing research studies major events or catastrophes. The results open new opportunities for supply chain managers to reduce weather dependence and improve profitability.

Keywords Weather, Demand variation, Seasonal products, Supply chain management and performance

Paper type Research paper

1. Introduction

Due to their multi-level structures, many events can disturb supply chain performance in several ways. Unplanned events may affect value and product processes, assets and infrastructures, inter-organizational networks through man-made and environmental causes (Peck, 2005). The more complex the supply chain is the more vulnerable it is to risks emerging from the supply and demand side, as well as catastrophic risks like natural hazards (Wagner and Bode, 2006). Major events or catastrophes make the news headlines and have significant consequences on societies, businesses and individuals. Along with the most memorable events, there have been severe but less extent incidents, such as European heat waves in 2003 and 2006, which were devastating and destroyed crops and affected businesses and their value chains. These and



similar events surely had a negative impact on supply chains, yet for some reason they seem to be within normality in their occurrence, and companies and people are expected to cope with them.

However, everyday weather-related fluctuations also affect supply chain performance in different business and supply chain contexts. Although these weather fluctuations, and even severe local weather conditions, occur frequently, most existing research focuses on rare major weather events or catastrophes and the vulnerability of supply chain operations. To fill this gap in the existing research, we focus in this paper on normal daily fluctuations in weather conditions by studying the following research question:

RQ1. Do everyday changes in weather conditions have an effect on demand, and in turn on supply chain performance measured by delivery punctuality?

Our research is based on analysing real transactional business data that covers over a decade of operations and supply chain-related transactions for seven different business units. These individual units are owned by a publicly listed brand holding company delivering well-known sporting equipment and goods to customers worldwide. The business units differ in operational strategy, product variety, country of origin, demand variability (seasonality) and predictability. The business data will be matched with daily location-based weather conditions (temperature, quantity of snow/rain, length of sunshine per day).

Our research methods include using a generalized linear model (GLM) to explain changes in demand and the delay in delivery based on local weather conditions, and then analysing the dependence between these using a generalized additive model and a copula approach. Our results show that, in addition to the order volume, customer locations and weather conditions affect supply chain performance. Further, this relationship is not constant but it depends on the season and weather conditions when resort locations are considered. The results can be used in many ways for improving supply chain performance. An example of this would be to remove bias caused by weather variables to measure the true supply chain performance of the company, to prepare the supply chain for higher volumes or variations in demand based on weather forecasts, or even to re-engineer supply chains which are less weather dependent.

Our study deals with weather events that are considered as normal in variation and which do not make the major headlines. We do not study floods, earthquakes and tsunamis, but the variation in performance and non-resilience in supply chains faced with normal fluctuations in weather conditions. We study orders made by business customers during business days, for example, retailers or ski rental companies, and delivery efficiency to them. Thus, end customer behaviour is beyond our scope. Since these normal changes in weather conditions do not have a direct effect on movement of goods, efficiency in factories, or logistics, we focus on weather at the customer's geographical locations.

We start by reviewing the literature on supply chain performance and vulnerability with a special focus on weather and its influence on performance. We seek to find gaps in the body-of-knowledge to form our detailed research hypotheses. The applied statistical methodology is then explained along with the description of the data used for the analysis. This is followed by a detailed analysis section after which the results are discussed from theoretical and practical points of view. We will also assess the internal and external validity of the work done. Finally, conclusions are drawn and avenues for future research are presented.

2. Literature review on weather affecting operations and supply chain management

Disruptions and process variance lead to poor supply chain performance. There is ample literature including case studies, models and surveys on the costly consequences of dysfunctional supply chains. In their seminal study, Hendricks and Singhal (2003) show that companies reporting problems in sourcing and delivery, product quality, etc., are associated with an abnormal decrease in shareholder value by 10.28 per cent. These problems refer to a vast multitude of incidents embedded in modern supply chains and researchers have classified and modelled these causes for poor supply chain performance in many ways.

For more than a decade research on supply chain risk, vulnerability and security has flourished and become a discipline of its own. Motivation for this research has increased ever since the 9/11 terrorist attacks (Sheffi, 2001), various epidemics (Giunipero and Eltantawy, 2004) and natural hazards and other drastic disruptions in the business environment (Kleindorfer and Saad, 2005; Sheffi, 2005; Sheffi and Rice, 2005). In their thorough analysis on supply chain disruptions, vulnerability and mitigation, Stecke and Kumar (2009) show that incidents negatively affecting supply chain performance have increased over time. These incidents include natural catastrophes, terrorist attacks, social unrest, major accidents and other mishaps that affect the increasingly complex and physically longer supply chains with increased numbers of exposure points. With regards to weather they hint that advanced companies do take forecasts into account when planning material flows. In their thorough study on supply resiliency, which results in a comprehensive framework, Blackhurst *et al.* (2011) treat weather as an extreme disturbance. Similarly, most of the literature on supply chain vulnerability concerns mainly major events and weather is only referred to in an extreme context.

Christopher and Lee (2004) emphasize the role of information sharing when mitigating supply chain risk while at the same time building confidence among the partners in the chain. In a similar vein, in their survey of supply chain professionals, Craighead *et al.* (2007) find that early warning systems and rapid distribution of information to various players in the supply chain are vital to prevent and prepare for potentially hazardous events. These warning systems should also include information on changes in weather, that is, should a ship be a day late or a truck a few hours off schedule. Delays in sharing information also imply delays in corrective measures, as demonstrated in the famous mobile phone industry case in which a storm triggered a fire at a critical component supplier's warehouse paralyzing the industry for several weeks (Norrman and Jansson, 2004; Latour, 2001). Although with no direct reference to weather, Manuj *et al.* (2014) indicate that postponement and reasonable speculation in supply chains may fall short of mitigating operational supply chain risk. They emphasize that it is vital to understand the total cost and system implications related to the importance of a stable and reliable supply base, the nature of demand variability, and the cost of finished goods inventory are reviewed first.

The impact of weather on productivity has been studied more widely than its impact on supply chain performance, especially in construction and agricultural industries. Thomas *et al.* (1999) quantified the effect of weather on a construction site and found significant losses in productivity because of snow (41 per cent) and cold temperatures (32 per cent). Seasonal industries have also examined the impact of weather. In their study on the construction industry, Rojas and Aramvareekul (2003) conclude surprisingly that external factors, notably weather and temperature, which are often cited as a major cause for reduced productivity, are considered to be one of the least relevant productivity drivers.

As Van der Vorst *et al.* (1998) show, even if average consumer demand is known there are always variations due to weather changes and changing consumer preference. The traditional way to respond to weather-induced fluctuations is by keeping inventory. From perishable goods to services, keeping inventory and reactive capacity are important as they are the main means to maintaining service level (Chopra and Lariviere, 2005). Further, Van der Vorst and Beulens (2002) study supply chain uncertainties through three case studies that were also vulnerable to weather. They indicate that weather plays a variation generating role both in up- and downstream supply chains, especially when agricultural and perishable products are concerned. In general, in food chains weather causes variation both in demand and supply and this should be taken into account. There are some indications that seasonal production, products and services tend to be more vulnerable to changes in weather (Costantino *et al.*, 2013), which partially explains the use of production smoothing and reactive capacity.

Clark and Hammond (1997) study the retail industry and the impact of electronic commerce to speed up reordering processes. They point out that efficient service also requires taking into account regional weather conditions (e.g. higher inventories are needed in Maine than in California due to snow and hurricanes) and that this is possible due to shorter replenishment cycles and information sharing. Following similar reasoning, Sheffi and Rice (2005) mention a company-specific weather service at a global parcel carrier incorporating weather fluctuations in their daily operational planning, which enables them to rapidly reroute to maintain the service level.

Aviv (2001) presents and formalizes the concept of collaborative forecasting as a means of improving supply chain performance. Even though the model is abstract, in a well concerted situation with continuously updated situational information, it could also take into account changes in weather. Chaharsooghi and Heydari (2010) study ways in which mean and variance in lead time affect supply chain performance. As lead time at each echelon of the supply chain plays a major role in the overall performance of the whole chain, they analyse whether the focus should be on the reduction of the mean or the variance. They show that focus on variance has a greater impact on supply chain and inventory performance, yet to tame the bullwhip effect focusing on reducing the lead time mean is more important. Ways to manage weather-induced variation are limited. However, as indicated earlier, being prepared via inventory and reactive capacity is one way to handle it. One could also prepare for fluctuations caused by weather through planning that extends beyond organizational boundaries.

Various financial instruments like rebates and derivatives also provide companies with a means of protecting themselves against disruptions and problems caused by weather. The development of weather derivatives represents one of the recent trends towards the convergence of insurance and finance (Brockett *et al.*, 2005). Chen and Yano (2010) show that the use of price fluctuations and hedging against bad weather, by using weather derivatives, could improve supply chain performance in weather intensive seasonal products. These tools aim to share supply chain risk along the downstream players of the supply chain. These and other statistical methods related to risk management have been applied to the supply chain context. These instruments are relatively new and mainly concern certain industries and markets. They are beyond the scope of our research, although there could be further uses for them in highly seasonal and weather dependent businesses.

There are many studies on the effect of the weather on consumer behaviour and demand. Bahng and Kincade (2012) study women's business wear and show strong evidence that fluctuations in temperature can impact sales of seasonal garments.

During sales periods when drastic temperature changes occur, more seasonal garments are sold. However, the temperature changes from day to day or week to week do not affect the number of garments sold for the whole season. They also show that fluctuations depend on the fabric and design.

Murray *et al.* (2010) study how weather affects consumer spending and the underlying psychological phenomenon. They especially focus on the sunlight effect. The authors recognize, based on the literature, three general categories: first, bad weather reduces people's willingness to go shopping; second, the weather has a direct effect on some products, such as ice-cream; and finally, the weather can affect the consumers' "internal states". The study contained sales and weather data for a period of six years. They find that an increase in sunlight tends to increase consumer spending but this effect also depends on temperature: at lower temperatures the effect is positive but negative when temperatures are high.

In agriculture, Behe *et al.* (2012) study the influence of weather on the sale of different plants (vegetables, flowers, etc.). Their conclusion is that the weather has an impact but it is weaker than that of the weekday, region, or month. Other examples of the weather effect on consumer decision making is a study by Busse *et al.* (2015) on how the weather conditions affect car sales. They find that the weather on the purchase day has a significant impact on the sales of convertibles and 4 × 4 vehicles. Bertrand *et al.* (2015) study how unseasonal weather affects apparel sales and how companies can deal with this risk. They use a linear regression model to estimate the impact of temperature differences on sales volumes in the apparel retail business.

There are also several studies on effects of external events such as weather on stock market behaviour. Levy and Galili (2008) study how cloudiness affects people's mood and their stock market transactions. They find differences among investor groups relative to how the weather affects their decision making. They conclude that, in cloudy weather, men, lower income and young people buy more stocks than other groups. In the same spirit, Lu and Chou (2012) study weather effects and stock index returns. Their conclusion is that weather can affect trading activities but not returns. Finally, an example of the influence of weather on the electricity market can be found in Huurman *et al.* (2012) who study the weather premium in the electricity market. They show that using the next day weather forecast clearly improves electricity price predictions.

As the literature review above illustrates, research into the effects of weather on supply chain performance is limited to overall classification of issues making supply chains vulnerable and ways in which major disruptions impact global supply chains. Weather plays a role in the operational planning of transportation, retail and seasonal businesses. However, the impact of fluctuations in temperature and sunshine duration on supply chain performance has not been systematically studied. Few case studies show that advanced companies can adjust and react to changes in weather, and can actually gain a competitive advantage from it (Ishikawa and Nejo, 1998). Common knowledge and some of the research shows that weather is blamed for poor performance, although this has not really been proven and the evidence is based on limited case-based surveys and anecdotal evidence.

3. Research hypotheses, data and methodology

Based on the existing literature, academics have paid limited research interest to weather and its impact on supply chain performance. On the one hand strong evidence supports the fact that weather affects demand and consumer shopping behaviour (Murray *et al.*, 2010; Parsons, 2001), and on the other hand, fluctuations in demand

affect supply chain performance (e.g. Beamon, 1999). But the overall chain of reasoning from fluctuations in weather affecting demand variation leading to variance in supply chain performance has been less studied. Although this effect is observed in some cases, it is not possible to establish such a general straight hypothesis. Therefore, this research aims to fill the gap in the current supply chain body-of-knowledge by studying ways in which weather affects demand in different market locations and different customer and product segments.

3.1 Building up detailed research hypotheses

By using longitudinal data covering about a decade of operations in several business units totalling over 300,000 deliveries of different sporting goods and matching these deliveries with daily weather data in different locations in Switzerland, we study the link between changes in weather, demand and supply chain punctuality. Depending on the product type, for example, summer or winter items, the change in temperature may have a different impact on demand, thus we analyse the direction of change in temperature. On the other hand, we assume that the greater the increase in order volumes measured in monetary value, the more the demand fluctuates, which in turn increases supply chain load and therefore a decrease in supply chain performance and punctuality.

The weather data holds daily information on the temperature differences to the long-term average, precipitation and sunshine duration relative to daily maximums. This data are location specific enabling us to divide our customers according to their geographical location into urban vs resort areas. We assume that resorts are more prone to demand fluctuations induced by changes in temperature, sunshine and snow fall. In urban areas meteorological changes may have less impact on demand than in resorts where people are more exposed to practicing activities related to the products studied. Our first hypothesis reads:

H1. Non-urban places, that is, resorts are significantly exposed to demand fluctuations induced by weather.

We then study seasonality. Some of the business units are seasonal (e.g. skiing equipment for winter sports), while others are by nature non-seasonal (e.g. sports instruments, certain apparel with relatively constant demand patterns). Seasonal here means that manufacturing, sales and actual deliveries take indistinct phases and overlap very little. These business units face the planning problematic depicted by the newsvendor model according to which one has to order goods in stock before facing actual and uncertain demand. For non-seasonal businesses the fluctuations in demand are smaller, and demand variations easier to anticipate, thus seasonal businesses are more prone to weather-induced disruptions in demand. Therefore we assume that seasonal businesses are more vulnerable to fluctuations in weather, and that they face more supply chain delays. Therefore the second hypothesis is:

H2. Supply chain performance (punctuality in terms of delay) for seasonal business is more vulnerable to supply chain load (order volume in local currency) than it is for non-seasonal business. Furthermore seasonal businesses and their supply chains react differently to fluctuations of weather than non-seasonal businesses.

The customers, meaning the retailers, speciality and sport shops, etc., vary hugely in sales volume and the amount of stock keeping units they hold. By sales rank we differentiate between small and large customers. We assume larger customers order more frequently and have more choice to offer and thus they hold larger buffers to

mitigate demand fluctuations. Therefore they also face better supply chain performance than smaller customers. On the other hand, as larger customers order larger volumes, changes in weather may increase their ordering volumes relatively more than the same weather changes for smaller customers. Thus larger customers may face more supply chain delays. On the other hand, smaller customers may be more reactive to changes in weather thus doing more rush orders and therefore they may be faced with more frequent supply chain delays. As the literature and our reasoning leads to differing views on how weather affects supply chain performance large and small customers, we write the third hypothesis in the following general form:

H3. Weather has a different impact on large and small customers' ordering behaviour, which in turn, affects their supply chain delay.

The three previous hypotheses state separately the effects of different variables on supply chain performance and on order volume. We then study the effects of these independent variables on the dependence between delay and order volume. If there is a relationship between high demand and delay, this association should be less significant in summer but it increases in resorts when the weather conditions deteriorate. This leads to our last hypothesis:

H4. The dependence between demand volume and supply chain performance is seasonal and is affected by weather in resort areas.

The statistical analysis shows that these hypotheses are not independent and therefore may not be studied separately. In Section 4, we precisely formulate our findings, looking at the significant interactions between the different variables and for different seasons of the year. Altogether the hypotheses make it possible to devise statistically justified sentences on the multidimensional relationships related to customer size, geographical location, seasonality, and weather fluctuations. The original motivation for the study stems from the supply chain management of the underlying case company, and their willingness to drill in to the very impact of weather on their supply chain performance. They have hands on knowledge and experience spanning over several years indicating that weather has an impact on their performance, yet they do not have a unanimous view on it and opinions tend to vary over the years along with financial results.

3.2 The case company and data

The case company, established in 1950, has a long history of commodity products. In the end of the 1980s a decision was made to transform the company entirely into a sporting goods company. The first sporting goods brand was acquired in 1989 followed by another four acquisitions over the next 15 years. At the time of the study, the company owned seven global business units providing customers with sports equipment for a range of summer and winter sports, indoor and outdoor sports, sports instruments as well as fitness equipment. We chose five brands for this study:

- (1) Salomon (seasonal): alpine skiing, cross-country skiing, snowboarding and trail running;
- (2) Atomic (seasonal): alpine skiing;
- (3) Wilson (non-seasonal): tennis, baseball, American football, golf, basketball, softball, badminton and squash;

-
- (4) Suunto (non-seasonal): sports precision instruments; and
 - (5) Precor (non-seasonal): premium-quality fitness equipment.

At the centre of our study are three constructs, namely, weather, supply chain load and operational performance. Weather is a construct formed by temperature and daily sunshine. Supply chain load is measured by order volumes in local currency and the operational performance of the supply chain is measured as a delay in days between actual and planned delivery date. The business contexts are urban vs resort customer residence areas and the delivery is destined for seasonal vs non-seasonal product families and small vs large customers. The data consist of 307,300 in-season orders in Switzerland and the corresponding delivery delays observed from March 2003 to December 2012. The in-season orders as opposed to pre-seasonal orders are driven by the situational constraints related to weather, overall demand and business expectations, and they are supposed to be delivered right away.

Products are manufactured or sourced based on sales plans for the following six to 12 months. Sales plans are based on historical sales, commercial insight and on open orders. Especially for seasonal products open pre-season orders provide a valuable source of information for making accurate sales forecasts. Products are distributed via regional distribution centres located in Austria, France and Germany. These distribution centres serve customers in all European countries, and are in some cases also replenished by regional distribution centres in Asia and the Americas. Manufacturing and sourcing volumes are defined using multi-level MRP, considering manufacturing capacity in its own and in supplier factories. The target of this production planning is to maintain a balance between high delivery precision and low inventory levels.

Switzerland is a mid-size country in terms of company sales. Products are typically delivered directly to shops. At the time of the data sample, own retail and e-commerce were still very small scale. There were no country-specific distribution facilities in Switzerland. The country was divided into twelve different areas classified as urban and holiday resorts. The urban areas were located around the cities of Basel, Bern, Geneva, Locarno, Lucern, Neuchâtel, Zurich, St-Gallen, while the studied holiday resorts are La Chaux-de-Fonds, Davos, Samedan (St. Moritz) and Sion. The data were collected from the enterprise resource management system that is used to manage each brand (for the data structure, see Table I). The data harmonization and integration process took place in phases as new brands were acquired.

The weather data were retrieved from the national weather database containing weather statistics on all weather stations in Switzerland. The data were matched with the chosen urban and resort areas mentioned above. The weather data holds continuous daily values on the key variables and a relevant geographical location related to main sales areas. Three of the resorts are in the mountain area, two of them being well-known ski resorts. The rest of the resorts are in main cities in Switzerland. We retrieved the temperature difference from the long-term average, as explained in the next section, and daily sunshine duration relative to possible daily maximum. The database for the analysis was built by merging the supply chain data and weather data based on the date and location. Each data item has the following form: customer, order ID, order time, order line value, delivery time, delay, weather station, customer size, postal code, product group, product ID, weather variables. For the weather variables we use moving average over the past five days.

Table I.
The structure of the supply chain transaction data

<i>Order line</i>	<i>Sales document</i>	<i>Product hierarchy</i>
Sales document number	Sales document number	Product hierarchy
Item	Sales doc type	Description
Material number	Sales organization	Brand
Quantity	Created by	<i>Material group</i>
Value	Sold-to party	Material group
Plant	Ship-to party	Description
Shipping point	<i>Material</i>	<i>Customer</i>
Created on	Material number	Customer number
Requested delivery date	Material description	Zip code
Realistic GI date	Product hierarchy	Country
Actual GI date	Material group	
Confirmed delivery date	Material type	
Customer pick-up date	Retail intro date	
Ready pack date		

Figure 1 shows the histogram for the logarithm of order line values. We use a GLM to explain the logarithm of the mean order line value, which we use to indicate the order volume. We use the following explanatory variables (covariates):

- *Place*: customer location either an urban city or a ski resort; and
- *Period/season* of the year: we define three main periods of a year that we call winter (from November to February), spring (from March to June) and summer (from July to October).

There are statistical reasons for having split the year in three main seasons: first the four formal seasons lead to too few data per season in the analysis and second, the decomposition which actually integrates formal “autumn” in Summer and Winter shows most significant effects among all possible yearly break downs by blocks of three months. We also checked seasonality in demand by adding the “day-of-year” covariate using a non-parametric form allowed by

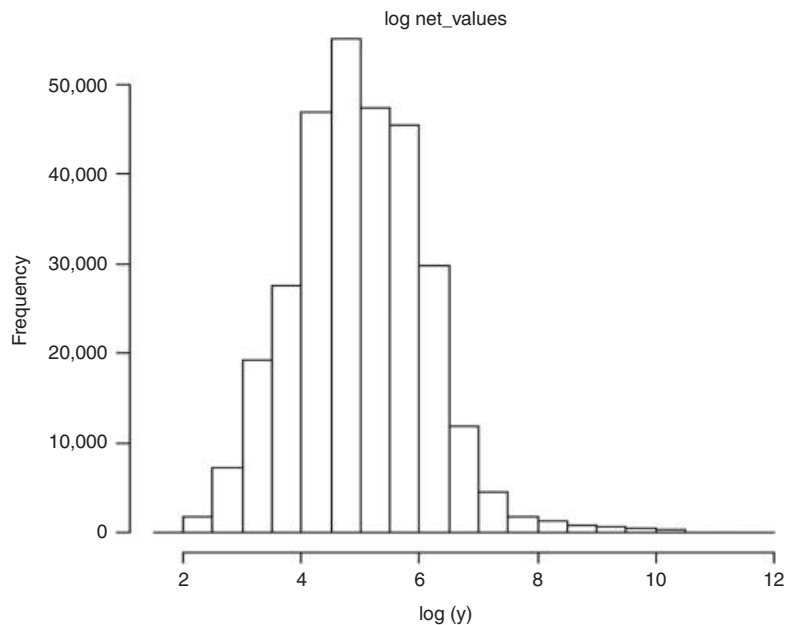


Figure 1.
Histogram of the logarithm of order volumes measured in monetary value

- generalized additive models. The effect exists but it is negligible under this non-parametric form. Thus, we decided to use a factor “season” instead.
- *Product type*: we differentiate between seasonal (Atomic, Salomon) and non-seasonal (Suunto, Precor, Wilson) products.
 - *Customer type*: small or large (determined by individual size in sales volume from the case company to customer): small customers are those below the median of the empirical individual size and large customers are above it. The orders among these groups are distributed as follows: small customers in urban areas 39 and 11 per cent in ski resorts; large customers in urban areas 37 per cent and in ski resorts 13 per cent.
 - *Weather variables*: we use moving average over the past five days of “temperature” (T). This is the temperature at 2 meters above ground which is a deviation from the daily maximum in relation to the “norm 6190” (norm 1961-1990). This variable is continuous and its histograms for the three periods are shown in Figure 2. The distributions are skewed, especially in spring, highlighting a warmer deviation from the norm (1961-1990). For the sake of interpretation, and in order to make it a factor, for each period of the year, we differentiate two levels of temperature: the low level called LowTemp for which the temperatures are below the median of all the temperatures observed for the related period of time, and HighTemp which represents the temperatures below the median. Similarly, we also use sunshine duration (high and low) and precipitation (high and low) as covariates but it turns out that these two are not significant to explain the demand.

3.3 Explaining weather effect for demand

We use a GLM to explain the mean value of an order line μ by the set of covariates listed above. The model reads $\mu = e^{(X\beta)}$, where X is the vector of covariates and β the vector of coefficients. The general purpose of the GLM model is to quantify the relationship between several predictor variables, their possible interactions and a dependent variable. The model explains the significance of predictor variables and it can also be used for forecasting values of the dependent variable. In our current case, the covariates are *place*, *period*, *product type*, *customer type* and *weather variables* and the dependent variable is the mean value of an order line which we use to indicate the *order volume*.

To find the most suitable model, several models including different covariates (X) are fitted and compared using the likelihood ratio statistic which is a standard test for nested models. We use the 5 per cent confidence level to retain significant covariates. Based on this procedure, the significant covariates are: *place* (urban vs resort), *period* (winter, spring or summer), *product type* (seasonal vs non-seasonal), *customer type* (small vs large) and *temperature variable* (LowTemp vs HighTemp). The interactions between the covariates are also significant. The goodness of fit of the model is assessed by standard diagnostics on residuals. The residuals are the error terms, that is, the differences between the fitted values (obtained from the model) and the observed values (the data). The resulting estimate coefficients are listed in Table AI in the three left columns. The table provides the coefficient estimates (β), their standard errors for all significant covariates and their interactions and the level of significance given by the stars. The R^2 of the model is about 21 per cent. The coefficients with the largest values correspond to the most significant factors in the model. The covariates, or their

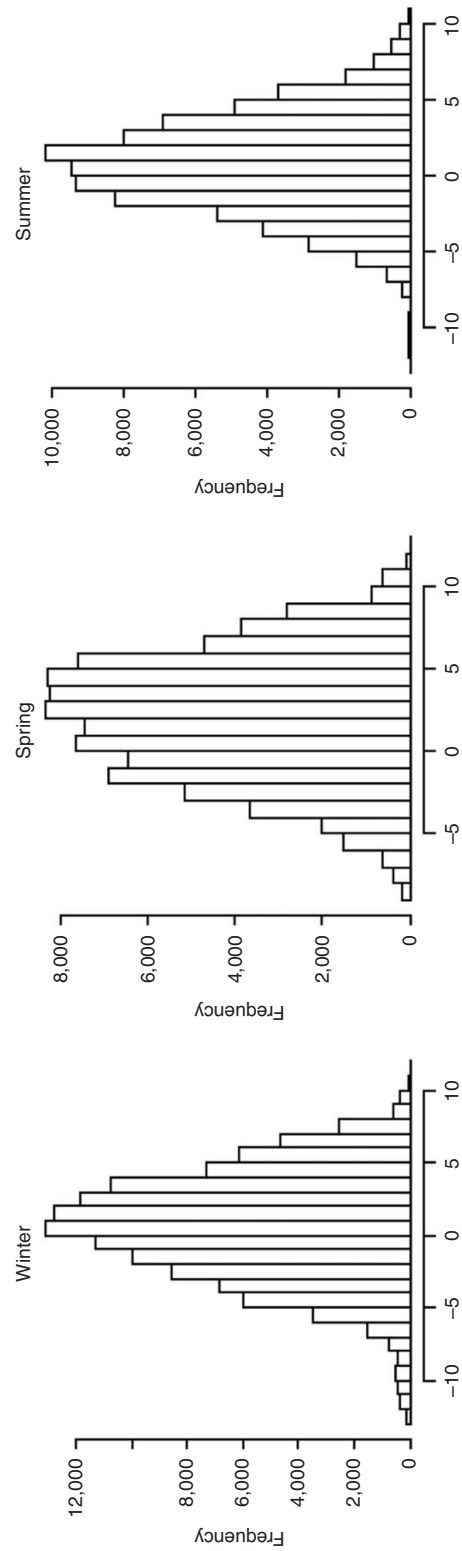


Figure 2.
Histograms of the temperature variable for winter (left), spring (middle) and summer (right)

interactions that do not appear, are set to 0. An example of using the table to calculate the mean value of order lines is the following: for a small customer in a resort ordering a seasonal product in winter during a low temperature period ($T < \text{median}$), the estimate mean of order lines is (in CHF):

$$e^{(4.93 + 0.89 - 0.03 + 0.28 + 0.08 + 0.07 - 0.24 + 0.03 - 0.61 - 0.20 - 0.1 - 0.01 + 0.25 + 0.07)} = 223.63.$$

Whereas the same characteristics (covariate levels) but for high temperature ($T > \text{median}$) gives an estimate mean order line value of CHF 208.51. The decrease in mean order line value when the temperature increases from LowTemp to HighTemp corresponds to one of the global trends discovered in our detailed analysis. If needed, one can provide confidence intervals for the estimators using the standard errors. One way to check these fitted values from the model is to compare them with the actual mean values observed in the data set. For the low temperature period, the actual mean was 222.19 and for the high temperature period 207.89. The predicted values from the model are therefore very close to the actual values.

The resulting GLM model can be used to predict the effect of weather variables on order volumes (with an uncertainty level) but it does not directly show the actual effects. Therefore, based on the significant coefficients of the model, we constructed tables (Table II, Section 4) using the actual data with rows and columns representing the levels of the different covariates. These tables clearly show the observed impact of the weather (temperature) on order volume.

	Weather (temperature) → order volumes				Order volumes → delay			
	Winter							
	Small customer		Large customer		Small customer		Large customer	
	Resort	Urban	Resort	Urban	Resort	Urban	Resort	Urban
Seasonal	↓ -0.24	-0.17	-0.16	↑ 0.29	↑ 0.78	↑ 0.83	↑ 1.61	↑ 2.31
Non-seasonal	↓ -0.21	-0.11	-0.07	0.03	-0.05	↓ -0.54	0.88	-0.11
	Spring							
	Small customer		Large customer		Small customer		Large customer	
	Resort	Urban	Resort	Urban	Resort	Urban	Resort	Urban
Seasonal	↓ -0.24	0.14	0.04	0.00	↑ 0.52	0.06	↑ 1.28	0.47
Non-seasonal	↓ -0.23	-0.06	0.04	↑ 0.25	↓ -0.27	↓ -0.62	↓ -0.29	↓ -0.41
	Summer							
	Small customer		Large customer		Small customer		Large Customer	
	Resort	Urban	Resort	Urban	Resort	Urban	Resort	Urban
Seasonal	-0.09	0.18	0.05	0.00	↑ 3.15	↑ 4.24	↑ 4.38	↑ 5.17
Non-seasonal	0.12	0.06	0.14	-0.13	↑ 0.81	↓ -0.46	↑ 0.41	-0.01

Notes: The numbers show the relative proportional difference for the given season. For instance, -24 per cent is the relative decrease in percent of order volumes for a small customer in a resort for a seasonal product in winter due to an increase of temperature. Only significant changes are shown by an arrow at the level of 5 per cent significance. Red means a decrease and blue an increase: the more intense the colour, the more significant the effect. It is possible to provide confidence intervals for the true proportions but they are not provided here to keep the paper focused

Table II. Increase in temperature effect on order volume (left blocks) and increase in order volume on delay (right blocks) during winter (top blocks), spring (middle blocks) and summer (lower blocks) seasons at different combinations of customer size, location and product type

3.4 Modelling delays by using weather information

We then study ways in which changes in weather affect supply chain performance, that is, the delivery delay in days. We assume that the delay follows a Poisson distribution with a rate that depends on the same covariates as used in Section 3.1, allowing for possible interactions. The used model is again a GLM model defined as follows: Denote by D the delay in days. Now D is supposed to follow a Poisson distribution with rate λ , such that $D \sim \text{Poisson}(\lambda)$. The GLM model reads as $\lambda = e^{(X\beta)}$, where X is the vector of covariates plus interactions as defined above and β the vector of coefficients. Several models are tested and compared. The selected model includes all the same covariates as in the previous sub-section: *place* (urban vs resort), *period* (winter, spring or summer), *product type* (seasonal vs non-seasonal), *customer type* (small vs large) and *weather variable* (lowTemp vs highTemp), plus their interactions. Table AI (three right columns) in the Appendix provides the estimate model. Figure 3 shows the histogram of the delay (left panel) and of the residuals of the model (right panel). The residuals vary around zero and are skewed, but much lower than the actual delay values. In that sense the model seems reasonable.

The interpretation of the Table and its usage is similar to the explanation provided in Section 3.1. For instance, for the same example of a small customer in a resort ordering a seasonal product in winter during a low temperature period ($T < \text{median}$), the estimate rate of delay days is:

$$e^{(2.44-0.47-0.24+0.06-0.15+0.07+0.25+0.2-0.04-0.27-0.02-0.09-0.07+0.45)} = 8.33.$$

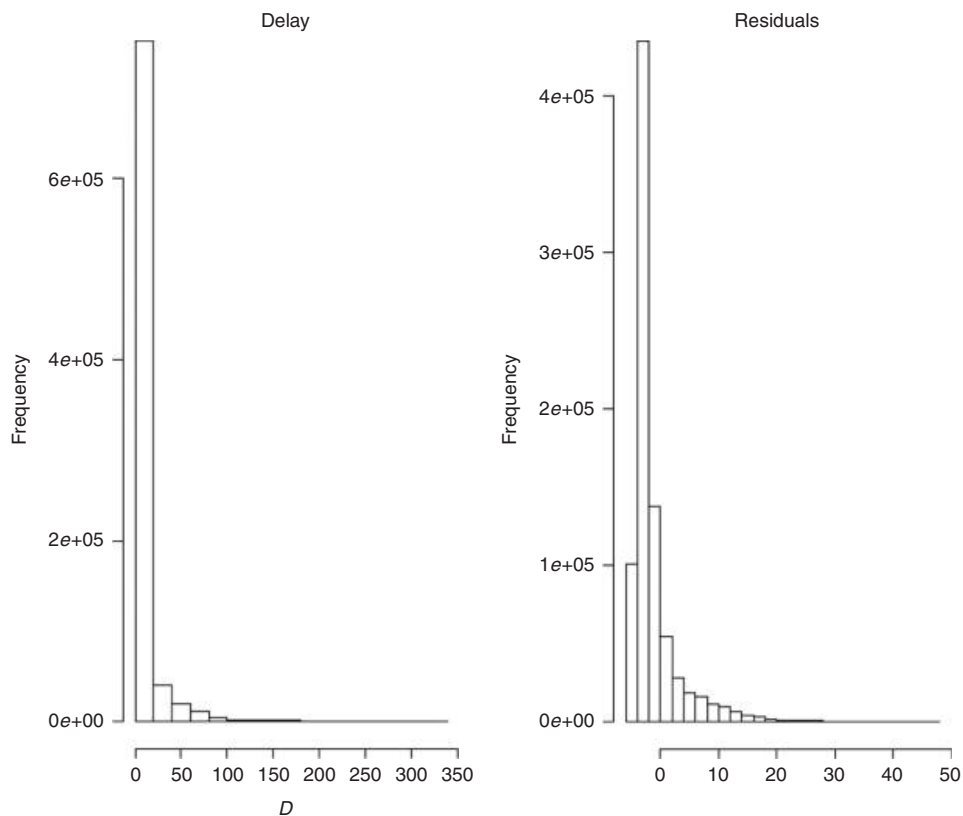


Figure 3. Histograms of the delay (left) and histogram of the residuals of the model for the delay (right panel)

The observed mean delay value is 8.19. Whereas the same characteristics (covariate levels), but for high temperature ($T > \text{median}$) gives an estimate of 7.61 days for the mean delay. The observed mean value is 7.51.

In conclusion, the example in Section 3.3 shows that when the temperature increases from LowTemp to HighTemp, the mean order line value significantly decreases for small customers in resorts ordering seasonal products in winter. Furthermore, when the temperature increases, the delay in days also significantly decreases. This is in accordance with our observations in the summary Table II in Section 4. More precisely, when the temperature increases, the order line value decreases and when the order line value decreases, the delay also decreases (again, for small customers in resorts ordering seasonal products during winter). Apart from the statistical results obtained from these two GLM models, the next section shows factual evidence of the variables effect based on the observed data.

4. Further results and hypothesis validity

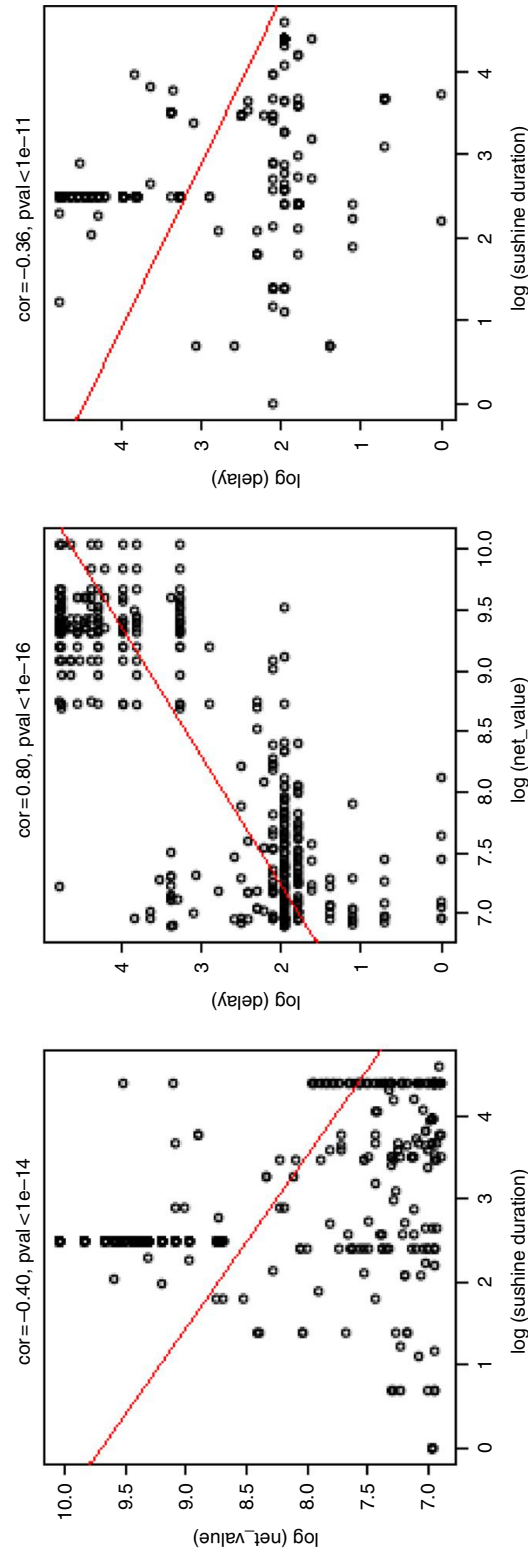
We start with a preliminary empirical analysis to determine whether changes in weather (temperature, precipitation) have an effect on supply chain load (delivery volumes measured in local currency), and whether this in turn has an effect on supply chain performance (punctuality). Figure 4 illustrates this in Zurich for a specific period of the year, product type and customer size. The left panel shows that sunshine has an effect on order volumes, while the middle panel shows that order volumes have an impact on delays and, finally, the right panel shows how supply chain delays are affected by sunshine.

We split our analysis in two: we start by studying how an increase in temperature affects order volumes and then we focus on how supply chain load affects performance. We divided the year into the three periods as described above and then we further drill in to urban vs resort areas and small vs large customers. Finally, we differentiate between seasonal and non-seasonal products. The schema in Table II summarizes the effect of increase in temperature on order volumes (left blocks) and the effect of increasing order volumes on delay (right blocks) in winter (top blocks), spring (middle blocks) and summer (lower blocks) for the combinations of customer size, location and product type.

The numbers are relative changes of proportion: the upper left value -24 per cent is the relative decrease in per cent of orders for small customers in resorts for seasonal products in winter due to the change of temperature from low to high. Only significant effects (every 5 per cent) are highlighted by arrows.

We can see from the table (left blocks) that an increase in temperature decreases order volumes in resorts both in winter and spring. In spring, larger customers in resorts increase their order volumes. In winter, changes in temperature have no effect on order volumes for small customers in urban areas, while for seasonal products rising temperatures increase order volumes for large urban customers. In spring, warmer temperatures increase sales for larger customers in urban areas. In winter, an increase in temperature reduces the order volumes of non-seasonal products at resorts. In all, for customers located in resorts, an increase in temperature always affects order volumes for seasonal and non-seasonal products (although sometimes non-significantly).

Table II (right blocks) shows that during the whole year, increasing order volumes significantly increase delays for seasonal products independently of the customer's size and location. For non-seasonal products ordered by large customers, winter causes delays at resorts when volumes increase. A similar analysis has been applied to order



Notes: The points in left panel correspond to the logarithm of order volumes (y -axis) in Zurich in winter (for big customer and seasonal product) against the logarithm of sunshine duration (moving average over the past five days, x -axis). The middle panel shows the corresponding delay values (y -axis) in logarithm against the log order volumes (x -axis). The right panel shows the log delay values against the log sunshine duration. The red line in each panel shows the linear dependence structure. The correlation appears above the panel and is significant dependence (p -value < 0.001) in each case

Figure 4.
Triangular effect

volumes using daily sunshine as the weather variable. Further empirical studies using other weather variables such as precipitation (rain or snow) confirm the hypothesis that weather has a significant effect on order volumes and thus it increases supply chain load. However, relative temperature is clearly the most significant weather variable in our case.

Regarding the hypotheses, the analysis shows that the highest temperatures significantly decrease order volume in a resort (especially in winter and spring). This means that a non-urban location or resort seems more vulnerable than an urban area to weather variability, which supports *H1*. The analysis also demonstrates that delay for seasonal products significantly increases with order volume, verifying the first part of *H2*. We also show that seasonal businesses react differently to fluctuations in weather than non-seasonal businesses, thus supporting the second part of *H2*. The same reasoning applies for the different (but not independently) effect on customer size, thus supporting *H3*.

So far, we have separately studied how weather fluctuations affect order volumes (measured in money) on one hand and delay on the other. However, it is then highly likely that the dependence structure between order volume and delay is itself significantly affected by weather fluctuation and other independent variables. To study this, we use the Kendall's τ as a rank correlation dependence measure between the order volume and delay. The used generalized additive model for the dependence measure is a copula-based approach developed in Vatter and Chavez-Demoulin (2015). In this approach, the generalized additive model framework (Hastie and Tibshirani, 1986) is extended to the conditional dependence structure; that is, it provides a very flexible model to explain the dependence measure (such as the Pearson correlation, Kendall's τ or Spearman's ρ) between two variables by independent variables. In our case, it is used to explain how the dependency between the order volume and the delay, as well as between the weather variables and the order volume, varies among seasons and weather conditions.

H4 stating that the dependence between demand volume and supply chain performance is seasonal and affected by weather in resort areas, is tested following the above copula approach and measuring the dependencies by using Kendall's τ . Our significant findings are first that the Kendall's τ varies yearly: in resort areas, order volume and delay are more dependent during winter than during summer. This is illustrated in Figure 5. The causal effect behind this finding is partly explained by the more difficult weather conditions in winter. This leads to the second point.

For seasonal products, the Kendall's τ between order volume and delay significantly increases with height of snow as shown in Figure 6 for Davos. For a large quantity of snow (above 1.5 meters), delay depends much more on demand than for reduced quantities of snow. To summarize, globally speaking, in non-urban places, delay is usually independent on demand but that changes once the weather conditions become difficult. These findings clearly support *H4*.

We also analysed the delay by the factor levels for urban place vs resort, seasonal vs non-seasonal, small vs large customer as well as their possible interactions and weather (measured by temperature) and order volumes. The aim was to detect crossed effects of weather and order volumes together on the different combinations of factor levels. The analysis is therefore complementary to the marginal effect documented above. From this further analysis we observe, for instance, that low-temperature and low-order volume lead to low delays for seasonal products. Respectively, high-temperature and high-order volume lead to proportionately high delays for seasonal products.

Figure 5.
Kendall's τ as
measure of
dependence between
delay and order
volume and its
yearly evolution in
resort places

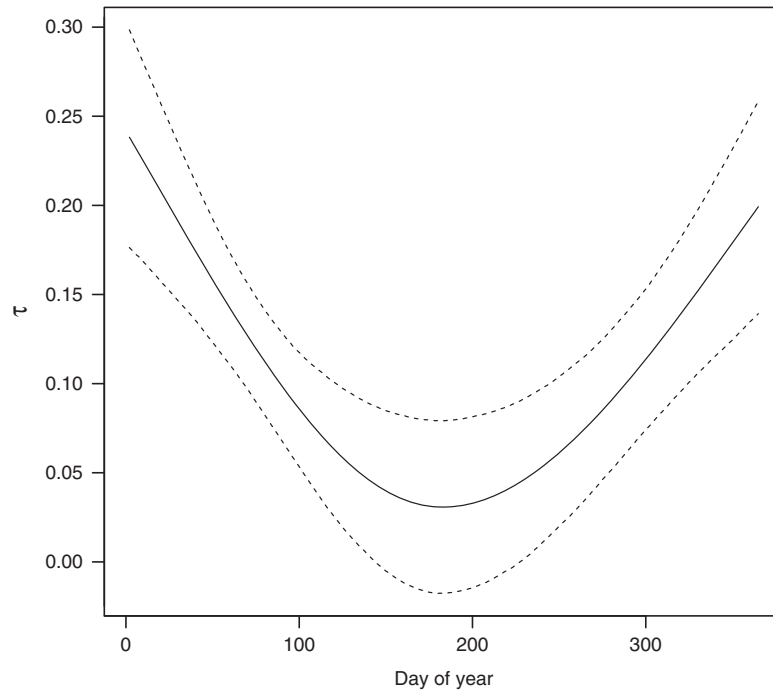
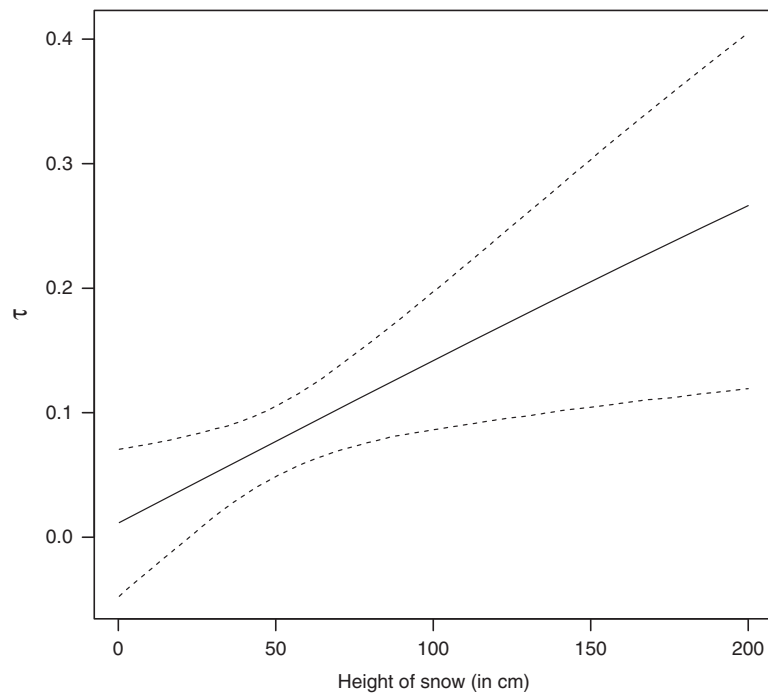
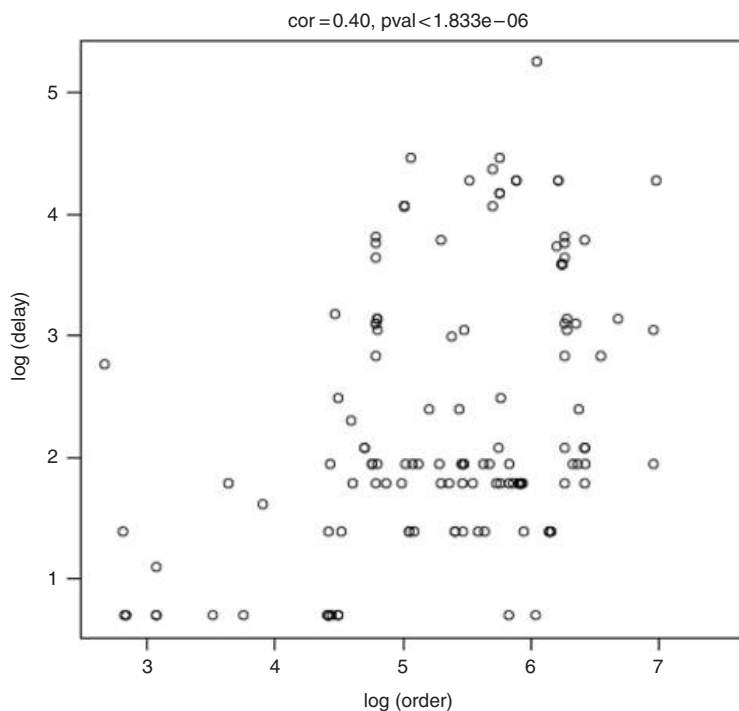


Figure 6.
Kendall's τ as
measure of
dependence between
delay and order
volume and its
evolution as function
of height of snow in
Davos, during winter



The marginal effect of increasing delays with increasing order volumes for seasonal products was observed. Marginally, we observed that high temperatures lead to globally lower order volumes of seasonal products (not for large urban customers). We also found that the increase in delays for seasonal products when order volumes increase is amplified by an increase in temperature. As an example, Figure 7 directly illustrates the effect of order volume on delay. The points are log delays against log



Notes: The correlation appears above the plot and highlights significant dependence ($p < 0.001$)

Weather and supply chain performance

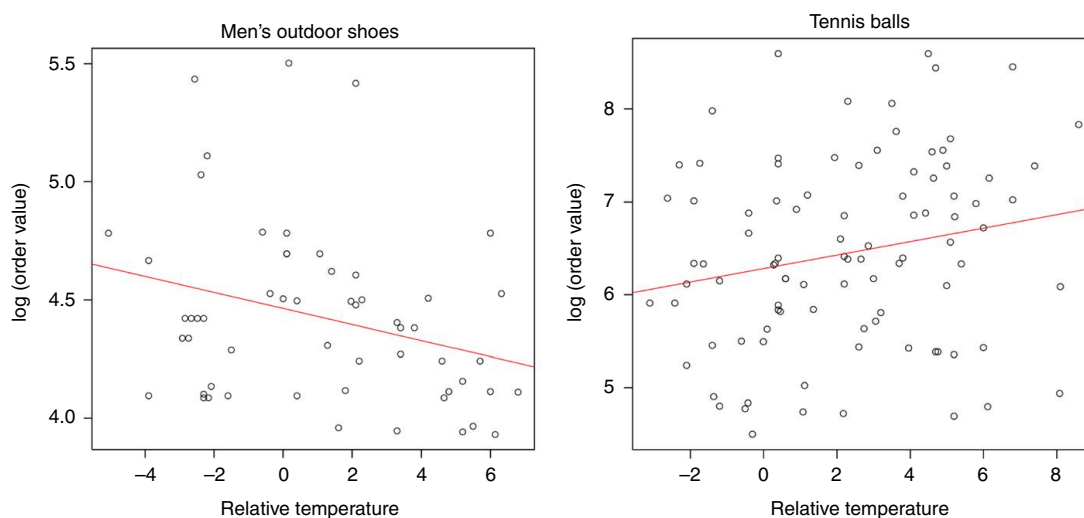
195

Figure 7.

The points are the log delay against log order volume for winter, seasonal product, ski resort, small customer and high temperature ($T > \text{median}$) in 2009

order volume in winter in a resort (Davos), for all seasonal products, small customers and temperature above the median for this period of time, in 2009. The correlation is significant and equal to 0.4.

The above documented statistically significant results emerge when analysing the data across different brands, yet product categories even within the same brand may behave differently to changes in weather. Figure 8 shows two categories, namely, men's outdoor shoes and tennis balls in Geneva during spring. These categories react in



Notes: Correlation for shoes -0.294 with p -value of 0.0277 and for tennis balls 0.206 and 0.0459 , respectively

Figure 8.

Two product families, men's outdoor shoes and tennis balls, and how they react differently to changes in temperature during a period from mid-march to mid-April in 2004-2012 in Geneva, Switzerland

different ways to changes in temperature. If the relative temperature increases, the demand for men's outdoor shoes will decrease, while the demand for tennis balls will increase. This may be trivial and obvious. However few companies extend and tailor their supply chain planning routines according to product category, not to mention incorporate weather forecasts in their planning routines. In most cases identical planning routines are applied to all product categories although they behave very differently. Based on the results of this research, the case company is planning to implement a pilot project where the sensitivity of a product category to changes in temperature is measured. This measure would then be used with a rolling forecast to anticipate demand fluctuation and at-hand inventory levels. Documenting these results would be a relevant topic for future research.

From the validity point of view our analysis is based on a single company, but on different business units with highly differing products. The constructs used are based on real supply chain transactions in the company's operations; orders. The weather data are also based on outputs from measurement units around Switzerland. As we use first-hand data, and no constructs or aggregates, the validity of the data can be considered sufficient. For internal validity the research should demonstrate that certain conditions lead to other conditions and to achieve this, multiple pieces of evidence from different sources are needed. The numerous analyses across business units, product groups and locations from which data were collected and analysed provide the study with the necessary internal validity. External validity of the research reflects whether the findings can be generalized beyond the immediate case. At least to a certain extent this has been achieved as the results are tested across and within cases, and the results do not conflict with previous research. The research concerns Switzerland, which limits the generalizability of the results. Additionally, the products concerned are mainly related to leisure, thus their demand may be more prone to weather inflicted variation.

5. Discussion and future research

We analysed how weather conditions and customer locations affect demand and especially supply chain performance. We used detailed transactional data on sales orders and deliveries combined with local weather data over a decade in Switzerland. We found two major results; first, weather at customer locations has a significant effect on order volumes and this effect differentiates the type of product and the location and size of the customer based on the season. Second, the order volume in monetary value, which corresponds to the load in the supply chain, has a significant effect on supply chain performance (punctuality in terms of delay). This means that weather affects supply chain performance.

The results can be used to estimate and explain the weather effect in supply chain performance. Further, our analysis shows that well-managed supply chains should be prepared for demand fluctuations due to weather changes. For example, monthly or seasonal weather forecasts could be used to estimate the punctuality of a supply chain and the supply chain should be prepared to handle large volume fluctuations caused by weather. Although medium range weather forecasts, that is, five to seven day forecasts, are still often inaccurate, they are continuously improving (Palmer and Weisheimer, 2012; Weisheimer *et al.*, 2011). Similar methods have been applied, for example, in predicting wind power capacity (Barbounis *et al.*, 2006). This suggestion is also supported by our finding that supply chains that are more used to dealing with weather fluctuations are even more robust when encountering variation in weather.

For the supply chain management of the case company the results show that weather fluctuations impact supply chain performance. Understanding this relationship is valuable both for improving sales forecast accuracy and for improving operational performance. To create an accurate sales forecast, historical sales is the most important starting point. This sales history will necessarily depend on weather conditions that will not repeat themselves the same as in past seasons. By using the model developed in this paper, the company can eliminate these weather effects to create a “normal weather condition” sales history as a base for forecasting. Second, it is possible to identify cases where weather conditions in past seasons affect subsequent season sales. For example, if the winter has been good, retailers tend to place more pre-orders for the next season – partly because their inventories are sold out, and partly because they are in a good mood.

Operational performance is largely affected by supply availability. An accurate long-term weather forecast would be of great help to further improve sales planning accuracy and make sure the right products are available in right quantities when needed. Unfortunately, weather forecasts do not forecast far enough into the future to cover sourcing and transportation lead time. However, product availability is not the only factor affecting delivery precision. Another important factor is distribution centre capacity. Picking, packing and shipping are manual operations that can be scaled by bringing in more personnel. Hence, understanding the relationship between weather, demand and operational performance can help the company utilize a ten-day weather forecast to adjust distribution centre capacity.

Finally, the company can re-engineer its supply chain to become less weather dependent. In fact, the company has announced that it has a strategic priority to “sustain profitability irrespective of weather conditions”. This is accomplished in three different ways. First, the overall product portfolio of the company consists of products for different seasons, for example, skiing and cycling, thus levelling the load while keeping utilization high and unit costs down throughout the year. This approach also enables economies of scope by sharing common resources like sales, marketing and distribution facilities across different product categories. Second, implications of the weather can be managed within each category by offering products for different weather conditions within the product category, for example, T-shirts and rain gear. Finally, every factory and supplier runs lean projects to reduce fixed costs, improve volume flexibility, and reduce lead time. As a result of this effort a large proportion of products can today be produced based on firm customer order, ensuring high customer service levels along with lower inventories.

The results of this study open new opportunities for the case company to reduce weather dependence and improve profitability. The fact that customers in different regions react differently to weather variations could be utilized for sales management. For example, if the weather is cold, the focus will be on selling to urban customers, because resort customers will buy in any case. If the weather is warm, the focus will be on resort customers. Results of the study could also be used for customer risk management, for example, as the resort customers are more influenced by weather, they should be offered less credit. Customers should also be educated to buy a balanced mix of different products: seasonal and non-seasonal, summer and winter. They will therefore be less vulnerable to fluctuations in weather.

The theoretical implications of this study are related to ways of managing supply chain variation. In general variation is divided into two different types: assignable or special cause variation and common cause or random variation. The former is generated by factors that can be identified and possibly managed – they stem from

external sources indicating that the process is statistically out of control. This type of variation could be caused by employees, different skill levels, mistakes in complying with the procedures, machine and truck breakdowns, etc. Common cause variation is inherent to the process – it is intrinsic to the process and will always be present. Weather is considered to be random variation, yet we know that advanced and agile supply chains can manage it partially through efficient inventory management, reacting faster and by better management of operational capacity. Our study takes this even further by showing that weather-induced variation could be treated more like special cause variation. Supply chains could be controlled better by understanding how weather affects different customers in size and their location and how product categories behave differently. As Christopher *et al.* (2004) argue, conventional organizational structures and forecast-driven supply chains are not adequate to meet the challenges of volatile and turbulent demand which typify fashion markets. Even though they call for agile organizational structures, one complemented with better understanding on the impacts of weather on supply chain performance would yield an even higher performance outcome. Taking into account weather in short-term supply chain planning makes it possible to serve the customer better, i.e. improved punctuality and order fulfilment. Theoretically, one could envisage that weather adjusted supply chain planning improves performance and therefore, through the increased volume, should also reduce prices and enlarge the choice customers have.

The findings reported open new avenues for future research to study how weather affects supply chain performance in different businesses and locations around the world, and especially, what management should do in practice to better manage their supply chains in alternating weather conditions. For the case company future research should include studying topics mentioned above, such as creating “normal weather condition” sales history, utilizing a ten-day weather forecast in planning, and reducing weather dependence. Additionally, further drilling down to understanding how different indoor/outdoor and summer/winter sport products behave, and how sentiment of the full season weather affects pre-orders for the next season, are also promising future research directions.

References

- Aviv, Y. (2001), “The effect of collaborative forecasting on supply chain performance”, *Management Science*, Vol. 47 No. 10, pp. 1326-1343.
- Bahng, Y. and Kincade, D.H. (2012), “The relationship between temperature and sales: sales data analysis of a retailer of branded women’s business wear”, *International Journal of Retail & Distribution Management*, Vol. 40 No. 6, pp. 410-426.
- Barbounis, T.G., Theocharis, J.B., Alexiadis, M.C. and Dokopoulos, P.S. (2006), “Long-term wind speed and power forecasting using local recurrent neural network models”, *Energy Conversion, IEEE Transactions on*, Vol. 21 No. 1, pp. 273-284.
- Beamon, B.M. (1999), “Measuring supply chain performance”, *International Journal of Operations & Production Management*, Vol. 19 No. 3, pp. 275-292.
- Behe, B.K., Getter, K.L. and Yue, C. (2012), “Should you blame the weather? The influence of weather parameters, month, and day of the week on spring herbaceous plant sales in the US Midwest”, *HortScience*, Vol. 47 No. 1, pp. 71-73.
- Bertrand, J.-L., Brusset, X. and Fortin, M. (2015), “Assessing and hedging the cost of unseasonal weather: case of the apparel sector”, *European Journal of Operational Research*, Vol. 244 No. 1, pp. 261-276.

- Blackhurst, J., Dunn, K.S. and Craighead, C.W. (2011), "An empirically derived framework of global supply resiliency", *Journal of Business Logistics*, Vol. 32 No. 4, pp. 374-391.
- Brockett, P.L., Wang, M. and Yang, C. (2005), "Weather derivatives and weather risk management", *Risk Management and Insurance Review*, Vol. 8 No. 1, pp. 127-140.
- Busse, M.R., Pope, D.G., Pope, J.C. and Silva-Risso, J. (2015), "The psychological effect of weather on car purchases", *The Quarterly Journal of Economics*, Vol. 130 No. 1, pp. 371-414.
- Chaharsooghi, S.K. and Heydari, J. (2010), "LT variance or LT mean reduction in supply chain management: which one has a higher impact on SC performance?", *International Journal of Production Economics*, Vol. 124 No. 2, pp. 475-481.
- Chen, F.Y. and Yano, C.A. (2010), "Improving supply chain performance and managing risk under weather-related demand uncertainty", *Management Science*, Vol. 56 No. 8, pp. 1380-1397.
- Chopra, S. and Lariviere, M.A. (2005), "Managing service inventory to improve performance", *MIT Sloan Management Review*, Vol. 47 No. 1, pp. 56-63.
- Christopher, M. and Lee, H. (2004), "Mitigating supply chain risk through improved confidence", *International Journal of Physical Distribution & Logistics Management*, Vol. 34 No. 5, pp. 388-396.
- Christopher, M., Lowson, R. and Peck, H. (2004), "Creating agile supply chains in the fashion industry", *International Journal of Retail & Distribution Management*, Vol. 32 No. 8, pp. 367-376.
- Clark, T.H. and Hammond, J.H. (1997), "Reengineering channel reordering processes to improve total supply-chain performance", *Production and Operations Management*, Vol. 6 No. 3, pp. 248-265.
- Costantino, F., Di Gravio, G., Shaban, A. and Tronci, M. (2013), "Exploring bullwhip effect and inventory stability in a seasonal supply chain", *International Journal of Engineering Business Management*, Vol. 5 No. 23, pp. 1-12.
- Craighead, C.W., Blackhurst, J., Rungtusanatham, M.J. and Handfield, R.B. (2007), "The severity of supply chain disruptions: design characteristics and mitigation capabilities", *Decision Sciences*, Vol. 38 No. 1, pp. 131-156.
- Giunipero, L.C. and Eltantawy, R.A. (2004), "Securing the upstream supply chain: a risk management approach", *International Journal of Physical Distribution & Logistics Management*, Vol. 34 No. 9, pp. 698-713.
- Hastie, T. and Tibshirani, R. (1986), "Generalized additive models", *Statistical Science*, Vol. 1 No. 3, pp. 297-310.
- Hendricks, K.B. and Singhal, V.R. (2003), "The effect of supply chain glitches on shareholder wealth", *Journal of Operations Management*, Vol. 21 No. 5, pp. 501-522.
- Huurman, C., Ravazzolo, F. and Zhou, C. (2012), "The power of weather", *Computational Statistics & Data Analysis*, Vol. 56 No. 11, pp. 3793-3807.
- Ishikawa, A. and Nejo, T. (1998), *The Success of 7-Eleven Japan: Discovering the Secrets of the World's Best-run Convenience Chain Stores*, World Scientific Publishing Co., Singapore.
- Kleindorfer, P.R. and Saad, G.H. (2005), "Managing disruption risks in supply chains", *Production and Operations Management*, Vol. 14 No. 1, pp. 53-68.
- Latour, A. (2001), "Trial by fire: a blaze in Albuquerque sets off major crisis for cell-phone giants – Nokia handles supply shock with aplomb as Ericsson of Sweden gets burned – was SISU the difference?", *Wall Street Journal*, 29 January, p. A1.
- Levy, O. and Galili, I. (2008), "Stock purchase and the weather: individual differences", *Journal of Economic Behavior & Organization*, Vol. 67 Nos 3-4, pp. 755-767.

- Lu, J. and Chou, R.K. (2012), "Does the weather have impacts on returns and trading activities in order-driven stock markets? Evidence from China", *Journal of Empirical Finance*, Vol. 19 No. 1, pp. 79-93.
- Manuj, I., Esper, T.L. and Stank, T.P. (2014), "Supply chain risk management approaches under different conditions of risk", *Journal of Business Logistics*, Vol. 35 No. 3, pp. 241-258.
- Murray, K.B., Di Muro, F., Finn, A. and Popkowski Leszczyc, P. (2010), "The effect of weather on consumer spending", *Journal of Retailing and Consumer Services*, Vol. 17 No. 6, pp. 512-520.
- Norrman, A. and Jansson, U. (2004), "Ericsson's proactive supply chain risk management approach after a serious sub-supplier accident", *International Journal of Physical Distribution & Logistics Management*, Vol. 34 No. 5, pp. 434-456.
- Palmer, T.N. and Weisheimer, A. (2012), "On the reliability of seasonal forecasts", ECMWF Seminar on Seasonal Prediction: Science and Applications, Reading, 3-7 September, pp. 185-194.
- Parsons, A.G. (2001), "The association between daily weather and daily shopping patterns", *Australasian Marketing Journal (AMJ)*, Vol. 9 No. 2, pp. 78-84.
- Peck, H. (2005), "Drivers of supply chain vulnerability: an integrated framework", *International Journal of Physical Distribution & Logistics Management*, Vol. 35 No. 4, pp. 210-232.
- Rojas, E.M. and Aramvareekul, P. (2003), "Labor productivity drivers and opportunities in the construction industry", *Journal of Management in Engineering*, Vol. 19 No. 2, pp. 78-82.
- Sheffi, Y. (2001), "Supply chain management under the threat of international terrorism", *The International Journal of Logistics Management*, Vol. 12 No. 2, pp. 1-11.
- Sheffi, Y. (2005), *The Resilient Enterprise: Overcoming Vulnerability for Competitive Advantage*, MIT Press Books, Cambridge, MA.
- Sheffi, Y. and Rice, J. (2005), "A supply chain view of the resilient enterprise", *MIT Sloan Management Review*, Vol. 47 No. 1.
- Stecke, K.E. and Kumar, S. (2009), "Sources of supply chain disruptions, factors that breed vulnerability, and mitigating strategies", *Journal of Marketing Channels*, Vol. 16 No. 3, pp. 193-226.
- Thomas, H.R., Riley, D.R. and Sanvido, V.E. (1999), "Loss of labor productivity due to delivery methods and weather", *Journal of Construction Engineering and Management*, Vol. 125 No. 1, pp. 39-46.
- Van der Vorst, J.G. and Beulens, A.J. (2002), "Identifying sources of uncertainty to generate supply chain redesign strategies", *International Journal of Physical Distribution & Logistics Management*, Vol. 32 No. 6, pp. 409-430.
- Van der Vorst, J.G.A.J., Beulens, A.J., Wit, W.D. and Beek, P.V. (1998), "Supply chain management in food chains: improving performance by reducing uncertainty", *International Transactions in Operational Research*, Vol. 5 No. 6, pp. 487-499.
- Vatter, T. and Chavez-Demoulin, V. (2015), "Generalized additive models for conditional dependence structures", *Journal of Multivariate Analysis*, Vol. 141, pp. 147-167.
- Wagner, S.M. and Bode, C. (2006), "An empirical investigation into supply chain vulnerability", *Journal of Purchasing and Supply Management*, Vol. 12 No. 6, pp. 301-312.
- Weisheimer, A., Doblus-Reyes, F.J., Jung, T. and Palmer, T.N. (2011), "On the predictability of the extreme summer 2003 over Europe", *Geophysical Research Letters*, Vol. 38 No. 5, pp. 1-5.

Appendix

Weather and supply chain performance

201

Covariates	GLM for log order volume			GLM Poisson for delay		
	Coefficients	SE	Signif	Coefficients	SE	Signif
HighTemp	5.00	0.02	***	2.34	0.01	***
LowTemp	4.93	0.02	***	2.44	0.01	***
Urban	-0.33	0.02	***	-0.43	0.01	***
Seasonal	0.89	0.02	***	-0.47	0.01	***
Small customer	-0.03	0.03	***	-0.24	0.01	-
Summer	-0.10	0.03	***	0.24	0.01	***
Winter	0.28	0.03	***	0.06	0.01	***
LowTemp & Urban	0.06	0.03	***	-0.18	0.01	*
LowTemp & Seasonal	0.08	0.03	***	-0.15	0.01	*
Urban & Seasonal	0.30	0.02	***	0.54	0.01	***
LowTemp & Small Customer	0.07	0.04	***	0.07	0.01	*
Urban & Small Customer	0.08	0.03	***	0.63	0.01	**
Seasonal & Small Customer	-0.24	0.03	***	0.25	0.01	***
LowTemp & Summer	0.15	0.04	***	-0.35	0.01	***
LowTemp & Winter	0.03	0.04		0.00	0.01	
Urban & Summer	-0.50	0.03	***	-0.63	0.01	***
Urban & Winter	-0.08	0.03	***	-0.16	0.01	*
Seasonal & Summer	-0.29	0.03	***	0.58	0.01	***
Seasonal & Winter	-0.61	0.03	***	0.20	0.01	***
Small Customer & Summer	0.17	0.03	***	-0.04	0.01	***
Small Customer & Winter	-0.20	0.04	**	-0.04	0.01	***
LowTemp & Urban & Seasonal	-0.21	0.03	***	0.36	0.01	***
LowTemp & Urban & Small Customer	-0.08	0.04	*	0.02	0.01	-
LowTemp & Seasonal & Small Customer	-0.10	0.04	***	-0.27	0.01	*
Urban & Seasonal & Small Customer	-0.18	0.03	***	-0.72	0.01	***
LowTemp & Urban & Summer	-0.07	0.04	***	0.18	0.01	-
LowTemp & Urban & Winter	0.00	0.04	***	0.61	0.01	
LowTemp & Seasonal & Summer	-0.11	0.05	***	0.35	0.01	*
LowTemp & Seasonal & Winter	0.00	0.04		-0.02	0.01	
Urban & Seasonal & Summer	-0.35	0.04	***	-1.68	0.01	***
Urban & Seasonal & Winter	-0.07	0.03	***	-0.03	0.01	*
LowTemp & Small Customer & Summer	-0.22	0.05	***	0.21	0.01	***
LowTemp & Small Customer & Winter	-0.01	0.05	***	-0.09	0.02	
Urban & Small Customer & Summer	0.31	0.04	***	0.71	0.01	***
Urban & Small Customer & Winter	0.00	0.04	***	0.31	0.01	
Seasonal & Small Customer & Summer	-0.05	0.05	***	0.06	0.01	
Seasonal & Small Customer & Winter	0.25	0.04	***	-0.07	0.01	***
LowTemp & Urban & Seasonal & Small Customer	0.17	0.04	***	-0.20	0.01	***
LowTemp & Urban & Seasonal & Summer	0.17	0.05	***	0.24	0.02	**
LowTemp & Urban & Seasonal & Winter	0.01	0.05	***	-1.10	0.02	
LowTemp & Urban & Small Customer & Summer	0.18	0.05	***	-0.20	0.02	***
LowTemp & Urban & Small Customer & Winter	0.00	0.06	***	-0.46	0.02	

(continued)

Table AI. Estimate coefficients and standard errors for the GLM model of Section 3.1 (two left columns) and for the GLM Poisson of Section 3.2 (two right columns)

Covariates	GLM for log order volume			GLM Poisson for delay		
	Coefficients	SE	Signif	Coefficients	SE	Signif
LowTemp & Seasonal & Small Customer & Summer	0.18	0.07	***	-0.27	0.02	**
LowTemp & Seasonal & Small Customer & Winter	0.07	0.06	***	0.45	0.02	
Urban & Seasonal & Small Customer & Summer	0.71	0.05	***	1.42	0.02	***
Urban & Seasonal & Small Customer & Winter	0.28	0.05	*	-0.16	0.02	***
LowTemp & Urban & Seasonal & Small Customer & Summer	-0.20	0.08	***	0.16	0.02	**
LowTemp & Urban & Seasonal & Small Customer & Winter	0.01	0.07	***	0.89	0.02	

Notes: Spring, non-seasonal products, large customers and resorts are taken as reference variables (their value is set to 0). Significance codes: **p*-value < 0.05; ***p*-value < 0.01; ****p*-value < 0.001

Table AI.

Corresponding author

Tapio Niemi can be contacted at: tapio.niemi@unil.ch

For instructions on how to order reprints of this article, please visit our website:

www.emeraldgroupublishing.com/licensing/reprints.htm

Or contact us for further details: permissions@emeraldinsight.com

Chapter 3

Using weather data to improve demand forecasting for seasonal products

Using weather data to improve demand forecasting for seasonal products

Flora Babongo*

Department of Operations,
Faculty of Business and Economics,
Anthropole, University of Lausanne,
CH-1015 Lausanne, Switzerland
Email: flora.babongobosombo@unil.ch
*Corresponding author

Patrik Appelqvist

Amer Sports Corporation,
Mäkelänkatu 91, P.O. Box 130,
FI-00601 Helsinki, Finland
Email: patrik.appelqvist@amersports.com

Valérie Chavez-Demoulin,
Ari-Pekka Hameri and Tapio Niemi

Department of Operations,
Faculty of Business and Economics,
Anthropole, University of Lausanne,
CH-1015 Lausanne, Switzerland
Email: valerie.chavez@unil.ch
Email: ari-pekka.hameri@unil.ch
Email: tapio.niemi@unil.ch

Abstract: In seasonal business, manufacturers need to make major supply decisions up to a year before delivering products to retailers. Traditionally, they make those decisions based on sales forecasts that in turn are based on previous season's sales. In our research, we study whether demand forecasts for the upcoming season could be made more accurate by taking into account the weather of the previous sales season. We use a ten-year dataset of winter sports equipment (e.g. skis, boots, and snowboards) sales in Switzerland and Finland, linked with daily meteorological data, for developing and training a generalised additive model (GAM) to predict demand for the next season. Results show a forecasting error reduction of up to 45% when including meteorological data from the past season. In our case, the value of this reduction in the forecasting error corresponds to around 2% of total sales. The results contribute to the theory of stochastic inventory control by showing that taking into account external disturbances, in this case the fluctuations in weather, improves forecasting accuracy in situations where the lag between ordering and demand is around one year.

Keywords: demand forecasting; seasonal products; newsvendor model; generalised additive model; GAM.

Reference to this paper should be made as follows: Babongo, F., Appelqvist, P., Chavez-Demoulin, V., Hameri, A-P. and Niemi, T. (2018) ‘Using weather data to improve demand forecasting for seasonal products’, *Int. J. Services and Operations Management*, Vol. 31, No. 1, pp.53–76.

Biographical notes: Flora Babongo is currently a PhD student at University of Lausanne, in the Faculty of Business and Economics. She received a Bachelor’s degree in Economics from University of Lausanne and a Master’s degree in Statistics from University of Neuchâtel. Before starting her PhD she worked two years for the Swiss Federal Office of Statistics. She is interested in applying statistics in different fields.

Patrik Appelqvist is Senior Director of Supply Chain and Operations Development at Amer Sports Corporation, one of the leading sporting goods companies in the world. He holds a PhD in Operations Management from Aalto University (2005). At Amer Sports, he has been driving projects related to demand forecasting, sales and operations planning, performance measurement and lean manufacturing.

Valérie Chavez-Demoulin is a Professor of Statistics at HEC Lausanne, specialising in statistical methods for quantitative risk management in general, and the statistical modelling of extreme events in particular. More recent methodological work concerns conditional dependence structures modelling, non-parametric Bayesian models, dynamic extreme value theory models and extremes for non-stationary time series. She is member of the RiskLab, ETH, Zurich and is an elected member of the International Statistical Institute (ISI).

Ari-Pekka Hameri is a Full Professor of Operations Management at University of Lausanne, Switzerland. He has been involved with numerous international research and consulting projects dealing with production and supply chain management. He has published over 70 articles in international management journals concerning the management of production, projects and supply chains.

Tapio Niemi holds a PhD in Computer Science from the University of Tampere (2001). He worked at the University of Tampere as a Scientist (1997–2001) and as an Assistant Professor (2002–2004). Since 2004 he has worked in the Technology Programme of the Helsinki Institute of Physics (HIP) at CERN, and since 2012 also as a Senior Scientist in the Department of Operations at HEC – Lausanne, Switzerland, focusing on green computing, data analytics, and business intelligence.

1 Introduction

“Prediction is very difficult, especially about the future”. This quote by Niels Bohr, Nobel laureate in physics, was about his understanding of atomic structure and quantum mechanics. However, it also holds for forecasting of seasonal products. A vast body-of-knowledge has been developed to predict and forecast demand, consumer behaviour, supplier and operational performance to optimise processes in various business contexts. We focus on seasonal products. In our case, this means studying a special sub-category of business decisions that fall under the general framework of the Newsvendor problem and stochastic inventory theory (Porteus, 1990; Gerchak, 2016). To be precise, we study detailed and real supply chain and demand data on seasonal products

to learn how past weather conditions affect accuracy of demand forecasting. Data related to external events, such as weather conditions, are usually not used for this type of decision making process, especially when decisions are made several months before the demand arrives. By seasonal products we mean winter sport goods, which are ordered and manufactured over 8 months before they are sold to consumers.

All companies have to cope with external variation and many of them collect past operational data to better understand the dynamics of their business environment. One of the main issues in demand forecasting is the bullwhip effect (Lee et al., 2004; Forrester, 1961), which states that variability in the ordering pattern increases when moving upstream towards manufacturers and suppliers. This phenomenon, originating from time delays, forecasting errors, long lead times and gaming etc., has been observed in most supply chains. For example one of the bullwhip effects on inventory is the net inventory variance amplification (Ma et al., 2013). Companies can mitigate the effect of demand uncertainty on their overall profit through operational hedging, which can be achieved via choice of product assortment (Devinney and Stewart, 1988; Treanor et al., 2014), accurate and/or quick responses using more flexible production capacity (Fisher et al., 1997; Sting and Huchzermeier, 2014), delayed product differentiation (Lee and Tang, 1997), resource diversification and sharing (Van Mieghem, 2007), along with logistics technology such as electronic data interchange to support quick responses to real demand information, and the usual in-season and end-of-season markdowns.

Our research is based on analysing real transactional business data that covers a decade of order line data on alpine skis of two well-known brands. These brands are owned by a publicly listed sporting goods company delivering well-known sporting equipment and goods to customers worldwide. The business data will be matched with daily location based weather conditions such as temperature, quantity of snow/rain, and length of sunshine per day. Our study deals with weather events that are considered as normal in variation. We focus on orders made by business customers during business days or trade shows, for example, retailers or ski rental companies one year before the actual demand occurs. The weather effect is not straightforward, since we only use past weather data, not the weather conditions during the time when the purchasing decisions were made, but the weather of the previous season. Our assumption is that the weather during the past season affects the sales of the next season. The reasoning behind this assumption is threefold:

- 1 If the previous season was good for retailers due to good weather conditions, they will have less left-over inventory and therefore need to order more products for next season.
- 2 Exceptionally good winter conditions may influence consumers to buy more.
- 3 Finally, based on Kahneman's (2011) availability heuristics higher sales during the previous season can have a psychological effect causing retailers to be more optimistic when placing their orders for the next season.

The main purpose of our study is to explain how weather conditions and fluctuations in weather affect the accuracy of demand forecasting using advanced statistical methods. As the main result of the study, we show that demand accuracy improves when weather data is combined with business data by using generalised additive models (GAM). This information can be used, for instance, to integrate the impact of variability in weather with company internal data in the decision making process to better anticipate demand

volumes. Our detailed results show that using past weather data to complement past seasonal demand data reduces the forecasting error up to 45% when compared to traditional forecasting models without weather information. To obtain these results, we start with a literature review on weather related demand forecasting and its impact on supply chain performance. We then formulate our detailed research hypotheses to fill in the gaps in the existing body-of-knowledge. This is followed by the description of the data and the explanation of the applied statistical methodology used for the analysis. After that, the results are discussed. Finally, conclusions are drawn and avenues for future research are presented.

2 Literature review

The following literature review starts by reviewing research aiming to integrate external information, especially weather, into the forecasting routines. We then proceed to review studies on how weather affects supply chain performance and productivity in different business situations. After this, we present studies exploring how weather affects consumer behaviour. Finally, we look into research presenting ways in which companies may protect themselves against weather variation.

Different forecasting methods have been developed for seasonal and dynamic businesses. For example, Thomassey (2010) proposes forecasting methods based on fuzzy logic, neural networks and data mining to improve supply chain performance in the clothing industry. Tabrizi and Ghaderi (2016) provide a comparative study of autoregressive integrated moving average and local linear neuro-fuzzy models through a manufacturing company case study. Aburto and Weber (2007) compare several forecasting techniques and provide hybrid systems combining an auto-regressive integrated moving average and neural network and fuzzy logic to improve supply chain management. Taylor and Xiao (2010) extend the value chain view by showing that the manufacturer benefits from selling to a better forecasting retailer if and only if the retailer is already a good forecaster. More generally, the manufacturer tends to be positively/negatively affected by improved retailer forecasting capabilities if the product economics are lucrative/poor. By extending the supply chain view Aviv (2001) introduces the notion of collaborative forecasting, where forecasting information is centralised and continuously updated in the replenishment process, with regard to demand evolution, which is impacted by external factors such as weather conditions. Appelqvist et al. (2016) study how variations in weather affect demand and supply chain performance in sport goods using generalised linear models (GLMs).

Weather has been widely studied and it has a clear impact on productivity, the supply chain, consumer behaviour and demand in general. For example Wal-Mart reported in June 2005 that its inventory levels were higher than normal because of below-normal temperatures (Chen and Yano, 2010). In a similar vein Coca-Cola and Unilever observed lower sales of soft drinks and ice creams because of colder than normal summers (Kleiderman, 2004). Amini and Ghodsi (2016) analyse an integrated transportation and inventory problem in a two-stage supply chain. They find out that in each period there are weather extra inventories or back-ordered demands. For many products weather represents an important factor in demand. This applies especially to products with high seasonality and long lead times in sourcing and delivery, because these characteristics make the company financially more vulnerable (Costantino et al., 2013; Wagner and

Bode, 2006). Starr-McCluer (2000) uses monthly data on retail sales and weather data to find out that unusual weather has a modest but significant role in explaining monthly sales fluctuations, thus contributing to the bullwhip effect. However, on an aggregate level lagged effects due to weather tend to offset original fluctuations and therefore, as with quarterly time windows, the weather effect tends to even out.

Weather conditions have a clear impact on productivity and supply chain performance. In the construction industry, Thomas et al. (1999) analyse three structural steel erection projects and quantify the effects of weather on productivity. Significant losses of productivity are observed, caused by snow (41%) and by cold temperatures (32%). Costantino et al. (2013) studied demand seasonality and argue that in production of products and services the demand may stem from factors such as weather, which partially explains the use of production smoothing and reactive capacity. This means that weather may also trigger the beginning of the season, for example, a longer warm period earlier in spring may advance the beginning of the sales period, and inversely bad weather may delay the start of the season.

Aggregate demand and supply is the result of individual consumer behaviour and the interaction with each other in the economy. Consumers are consciously or unconsciously aware of the weather and they are not exempt from its impact. Van der Vorst et al. (1998) show that even if average consumer demand is known, there are variations due to weather changes and changing consumer preference. The traditional way to respond to weather induced fluctuations is to keep inventory. From perishable goods to services, keeping inventory and reactive capacity are important as they are the main means to maintain service level (Chopra and Lariviere, 2005). Van der Vorst and Beulens (2002) study supply chain uncertainties through three case studies that were vulnerable to weather. They indicate that weather plays a variation generating role both in up- and downstream supply chains, especially when agricultural and perishable products are concerned. In a more in-depth study on agricultural products, Behe et al. (2012) study the influence of weather on the sales of different plants (vegetables, flowers etc.). They conclude that weather has an impact, but that it is weaker than the effect of weekday, region or month.

Murray et al. (2010) study the effects of weather on consumer spending. They provide empirical evidence to explain how weather affects consumer decision making, and detail the psychological mechanism that underlies this phenomenon. The authors find that temperature, humidity, snow fall and, especially sunlight can affect retail sales. They mainly analyse the effects of sunlight which is mediated by a negative effect, meaning that as exposure to sunlight increases, the negative effect decreases and consumer spending tends to increase. Bahng and Kincade (2012) explore this further and identify a relationship between temperature and retail sales of seasonal garments. Even though they had limitations in the sample and the location of stores, they provide interesting results. By analysing women's business wear, the authors show strong evidence that fluctuations in temperature can impact sales of seasonal garments. During sales periods when drastic temperature changes occurred, more seasonal garments were sold. However, the temperature changes from day to day or week to week did not affect the number of garments sold for the whole season. Further, the fluctuations depend on the fabric and design. In a seemingly non-seasonal business Bertrand et al. (2015) show that weather affects apparel sales by using a linear regression model to estimate the impact of temperature differences on sales volumes in the apparel retail business. In the car industry, Busse et al. (2015) study how the weather conditions affect car sales. Their finding is that for the sales of convertibles and 4×4 vehicles, the weather of the purchase

day has a significant impact. Fashion and luxury items also set a special demand for the organisation to react on demand fluctuations as Christopher et al. (2004) argue that conventional organisational structures and forecast-driven supply chains are not adequate to meet the challenges of volatile and turbulent demand which typify fashion markets.

Integration of weather forecasts in operational planning is also widely used. Steinker and Hoberg (2015) incorporate weather data into the sales forecasting of one of the largest European fashion online retailers. Using weather forecasts they are able to improve the sales forecasting accuracy incrementally by 57% on summer weekends. These considerable improvements in forecast accuracy may have an important impact on logistics and warehousing operations. They also apply their empirical results to quantify the value of incorporating weather information into workforce planning within an order fulfilment centre and show how weather data can be leveraged to improve costs and performance. They estimate that a weather forecast improved sales plan reduces excess costs over a perfect information scenario in the range of 12.2% to 12.9% relative to the baseline model. Few case studies (Ishikawa and Nejo, 1998) show that advanced companies can adjust and react to changes in weather, and can actually gain a competitive advantage from incorporating weather forecasts in their operational plans.

In finance, Levy and Galili (2008) analyse the effects of cloud coverage on individuals' mood tendencies to buy and sell equity. They found that the effects of cloudiness are different among groups based on age, sex, and income level. The conclusion was that in cloudy weather men, lower income individuals, and young people buy more stocks than others. In a similar vein, Lu and Chou (2012) study how weather affects stock index returns. Their conclusion is that weather can impact trading activities, but not returns. Goetzmann et al. (2014) study and show that weather-based indicators of mood impact perceptions of mispricing and trading decisions of institutional investors. Apergis et al. (2016) empirically show that unusual deviations of weather variables from their monthly averages have a statistically significant effect on stock returns across global returns.

With regards to market pricing, Huurman et al. (2012) show that predicting the price of electricity can be significantly improved by complementing traditional price estimation models with next day weather forecasts. Various financial instruments, like rebates and derivatives, also provide companies with a means to protect against disruptions and problems caused by weather. The development of weather derivatives represents one of the recent trends toward the convergence of insurance and finance (Brockett et al., 2005; Cui and Swishchuk, 2015). Nicola (2015) analyse the impact of weather insurance on consumption and welfare gains of farmers, and she finds that weather insurance has the potential to provide large welfare gains, equivalent to a permanent increase in consumption of almost 17%. Chen and Yano (2010) show that by using weather derivatives the use of price fluctuations and hedging against bad weather could improve supply chain performance in weather intensive seasonal products. These tools aim to share supply chain risk along the downstream players of the supply chain (Singh and Acharya, 2015). These and other statistical methods related to risk management have been applied to the supply chain context. These instruments are relatively new and mainly concern certain industries and markets.

Table 1 Summary table of the literature review under the main themes: forecasting, productivity, consumer behaviour, forecasts in operational planning and finance

<i>Theme</i>	<i>Results/observations</i>	<i>References</i>
Forecasting	Research has focused on forecasting methods based on fuzzy logic, neural networks, data mining, GLMs and formulation of hybrid systems combining auto-regressive integrated moving average to improve supply chain performance. Advantages of good forecasts have been illustrated. Finally benefits of collaborative forecasting, where forecasting information is centralised and continuously updated in the replenishment process.	Thomassey (2010), Tabrizi and Ghaderi (2016), Aburto and Weber (2007), Taylor and Xiao (2010), Aviv (2001), Appelqvist et al. (2016)
Productivity	Research shows that weather affects the productivity via inventories and sales, and is an important factor in demand, especially for high seasonality or perishable products.	Chen and Yano (2010), Amini and Ghodsi (2016), Kleiderman (2004), Costantino et al. (2013), Wagner and Bode (2006), Starr-McCluer (2000), Thomas et al. (1999), Chopra and Lariviere (2005)
Consumer behaviour	The effects of weather on consumer behaviour are proven by analysing sales of different plants, seasonal and non-seasonal apparel, cars and fashion and luxury items. Empirical evidence to explain how weather affects consumer decision making is provided and the psychological mechanism that underlies this phenomenon is detailed.	Van der Vorst et al. (1998), Van der Vorst and Beulens (2002), Behe et al. (2012), Murray et al. (2010), Bahng and Kincade (2012), Bertrand et al. (2015), Busse et al. (2015), Christopher et al. (2004)
Forecasts in operational planning	The incorporation of weather data into sales forecasting leads to an increase in competitive advantage.	Steinker and Hoberg (2015), Ishikawa and Nejo (1998)
Finance	Literature focused on analysing the effects of weather on stock index returns, on tendencies to buy and sell equity, and on the incorporation of weather forecasting in price estimation models. Research also studied the use of price fluctuations and hedging against bad weather by using weather derivatives.	Levy and Galili (2008), Lu and Chou (2012), Goetzmann et al. (2014), Apergis et al. (2016), Huurman et al. (2012), Leon et al. (2015), Brockett et al. (2005), Cui and Swishchuk (2015), Nicola (2015), Chen and Yano (2010), Singh and Acharya (2015)

To summarise, weather has a significant impact on consumer behaviour which manifests itself in demand fluctuations and even in the timing of the seasons. This impact varies greatly between product and service categories (agricultural, standard, commodity, fashion or luxury) etc. and several companies show that incorporating past performance with weather data can be used to improve operational and supply chain performance. Most advanced companies in highly cyclical industries also use weather forecasts in their production and supply chain planning. Our focus, however, is on seasonal products that have long delays, (i.e., several months) between ordering and delivery. Concerning

seasonal businesses, current research mainly focuses on how companies mitigate weather induced demand fluctuations. Estimating demand for a sales season that is months away is still mainly based on the analysis of the demand information from the past sales seasons. The most recent season is most dominant when making sales estimates. This is the gap our research aims to fill. We study how demand forecast based on demand information from past sales seasons can be improved with weather data gathered during the same seasons, when the lag between ordering and delivery is more than half a year.

3 Motivation, research questions, data and methodology

3.1 Motivation and research questions

Pre-orders play an important role in a seasonal business, since they cover around 80% of total order value. Therefore, accurate demand forecasting is a key element for successful business and even small improvements in prediction accuracy have a direct effect on the profit of the company, namely on customer service, gross margins, and inventories.

While numerous companies incorporate short term, that is, 1–5 day, weather forecasts into their short-term operational and supply chain planning, our study extends this to seasonal products where the demand occurs months after the manufacturing orders are placed. By using real ordering and demand data on products sold in wintertime, we show that it is possible to make more accurate demand forecasts for the coming season by taking into account the weather of the previous sales season. This means that despite the long one year lag between sales seasons, weather adjusted forecasts are more accurate than the traditional forecasts based on historical demand only. This main objective transforms into the following three main research questions on seasonal demand forecasting when the lag between consecutive seasons is one year:

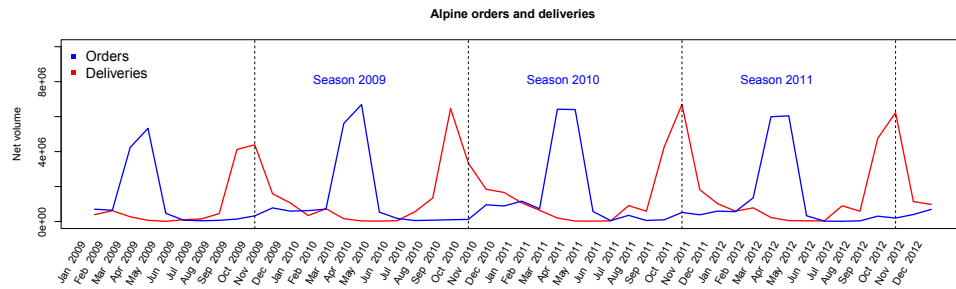
- How do weather conditions of the past sales season affect seasonal demand of the next season?
- How can past weather data be used to improve demand forecasting for seasonal products?
- How generalisable are the results when compared to other geographical areas?

3.2 Data

The case company, established in 1950, is one of the leading sporting goods companies in the world. The Amer Sports sales network covers 33 countries: the largest markets being the United States, Germany, France, Japan, Canada and Austria. Amer Sports sells its products to retail customers, which include sporting goods chains, specialty retailers, mass merchants, fitness clubs, and distributors. In 2015, Amer Sports net sales totaled EUR 2,534 million and currently the company owns several global brand business units providing customers with sports equipment for a range of summer and winter sports, indoor and outdoor sports, sports instruments as well as fitness equipment. Each brand has their distinctive operational strategies and organisations as the case company has acquired them over the past two decades.

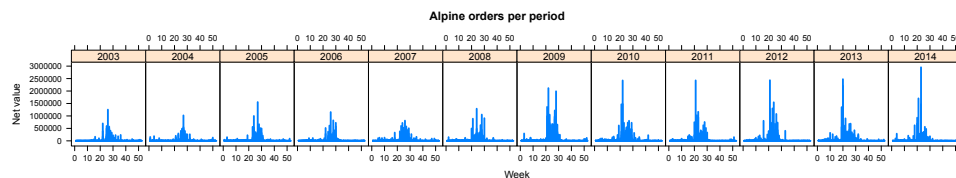
In this paper, we consider alpine ski products of two different brands, namely Atomic and Salomon. For these brands all decisions concerning production volumes have to be made well in advance, sometimes even eight months before the demand occurs (Figure 1). Normally, around 80% of the orders are made pre-season, the rest being ordered during the sales season. Taking pre-orders starts in early October and continues to the following spring while the deliveries of the products do not start until June of the following year. To handle this particularity, the ‘season’ spans from October to the end of September of the following year. This seasonal cycle, for example the 2009 season, covers the time between October 2009 and September 2010. This cycle repeats itself in a more or less similar form every year as can be seen in Figure 1. Based on this regularity, we assume that statistical methods can be applied to predict annual values for the order volume.

Figure 1 Orders (blue) and deliveries (red) for Alpine ski products for the years 2009–2012. Here the annual business cycle does not relate to the calendar year. This means that the 2009 season actually starts in October 2009 and ends in September 2010 (see online version for colours)



The company dataset contains more than 670,000 orders of winter sport goods, mainly alpine skies and their accessories issued by more than 1,000 customers between 2003 and 2014 (Figure 2). This data is matched with the daily weather data in different customer locations in Switzerland. The customers are business customers, mainly sport stores and wholesalers, not end consumers. The weather data holds daily information on the temperature differences to the long term average, daily maximum/minimum temperatures, precipitation, snow, humidity, pressure, wind and daily sunshine duration. Using these datasets, we study how weather affects demand and the incremental improvement gained by incorporating weather data with demand forecasting.

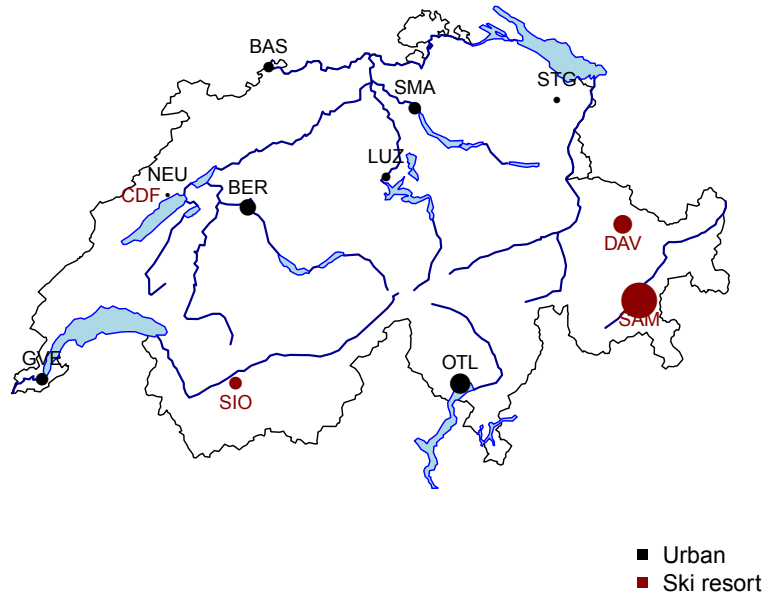
Figure 2 Weekly order volume distributions according to sales season, from 2003 to 2014 (see online version for colours)



3.3 Methodology

The weather dataset was retrieved from the Swiss national weather database containing weather statistics on all weather stations in Switzerland (Figure 3). Each ordering customer is assigned to the nearest weather station. In Figure 3 the size of the points is proportional to the associated total order volume in the area of the weather station. We apply a moving average with a lag of two days in order to smooth the weather variations. In other words, the quantity of snow used for day d is an average of three values: snow observed at day $d-2$, snow observed at day $d-1$ and snow observed at day d . To obtain the final dataset, we first merge company and weather data according to date and weather station. We then aggregate the monetary volume of orders in the area of each station. The latter manipulation is justified by the fact that aggregating more customers improves the relative forecasting performance up to a specific point (Sevlian and Rajagopal, 2014).

Figure 3 Swiss weather stations and main ski resort areas with their corresponding order volumes indicated by the size of the dot (see online version for colours)



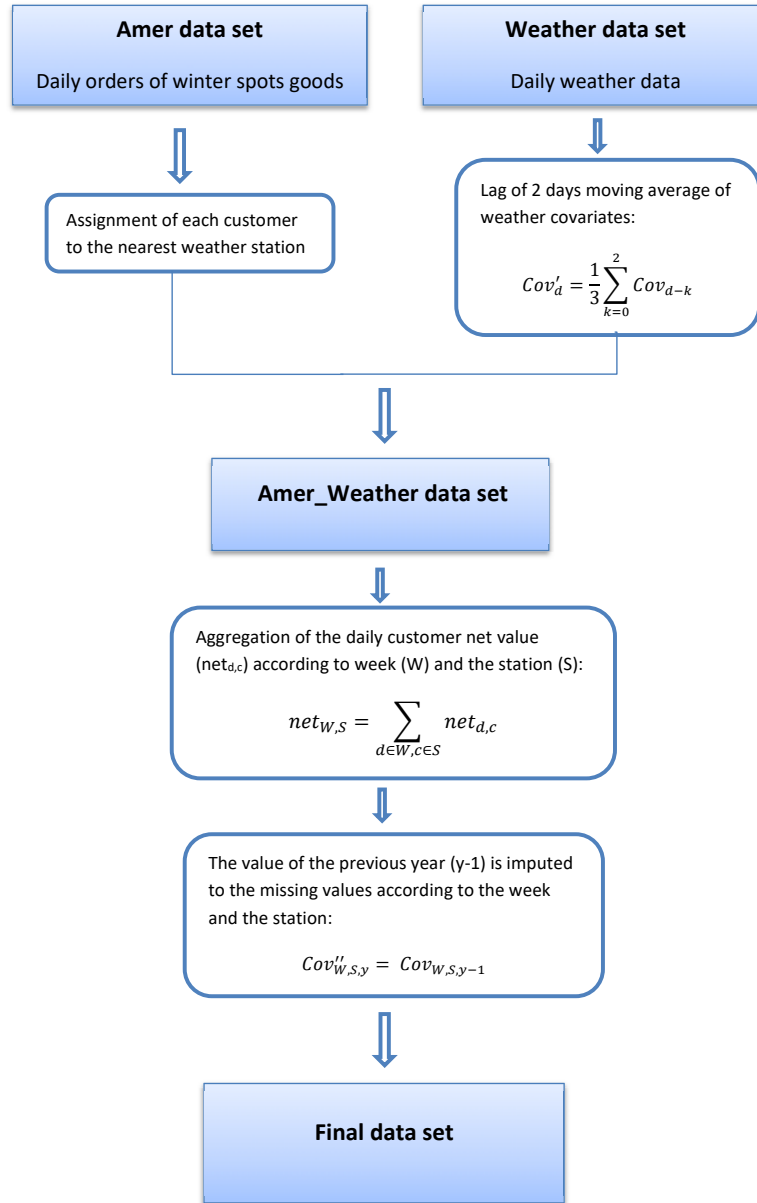
Since the actual day of the order is somehow ‘random’, meaning an order can be given late today or early tomorrow, and because weekends and other holidays occur differently each year, we work with aggregated weekly order volumes. This means that a customer is more likely to order in the same week each year rather than the same day each year. Hence, after aggregating the net values of orders according to the geographical location of customers, we sum them at the weekly level. For weather data, we apply the normal arithmetic mean function to obtain the weekly averages. We also noticed some missing values in weather variables, which correspond to less than 2% of the data. For these, we

imputed the value of the year before, according to the weather station and the week of the year. For example, if the quantity of snow is missing for weather station x in week t in 2008, then we impute the quantity of snow of the same weather station x and the same week t , but in 2007. We also dismissed the variables with the most missing values, thus the data used in the final prediction model contains only few missing values. The final dataset obtained after this procedure is described in Table 2 and all data processing steps are illustrated in Figure 4.

Table 2 The final merged dataset linking weather data with company data

<ul style="list-style-type: none"> • Order volume: in Swiss Francs (CHF) • Order week (week of sales season) • Season: seasonal cycle from October to September of following year • Weather season: spring, summer, autumn, winter • Station: the geographical location of the nearest weather station to the customer (Figure 3). They are divided into two groups: <ul style="list-style-type: none"> • Urban areas: Basel (BAS), Bern (BER), Geneva (GVA), Locarno (OTL), Luzern (LUZ), Neuchatel (NEU), Zurich (SMA), and St-Gallen (STG). • Ski resorts: La Chaux-de-Fonds (CDF), Davos (DAV), St. Moritz (SAM) and Sion (SIO). • Altitude: altitude of weather stations (m) • Snow: thickness of snow measured at 05:40 am (cm) • Fresh snow: thickness of fresh snow, sum of the day (24 hours), measured at 05:40 am (cm) • Pressure: atmospheric pressure at altitude of station, daily average (hPa) 	<ul style="list-style-type: none"> • Temperature_dv: temperature at 2 metres above ground, which is the deviation from the daily maximum in relation to the 'norm 6190' (norm 1961–1990), (°C) • Temperature_max: temperature at 2 metres above ground, daily maximum, (°C) • Temperature_min: temperature at 2 metres above ground, daily minimum, (°C) • Precipitation_1: sum of the daytime from 7am to 7pm, (mm) • Precipitation_2: sum of day (24 hours), (mm) • Humidity: relative air humidity at 2 metres above ground, daily average (%) • Sun_1: sunshine duration, relative to the daily maximum possible, (%) • Sun_2: sunshine duration, sum of the day, (min) • Wind: wind speed, daily average (m/s) • Gust: daily maximum (integration at 1 s) (m/s)
--	--

We use the same approach on similar datasets for Finland. The Finnish operations related dataset covers ten years of transactions also ranging from 2003 to 2014 and containing more than 240,000 order lines. Like with the Swiss case, the orders are aggregated according to the station and week, then merged with customer location specific weather data. The final dataset obtained contains the following variables: order volume, order date, station, snow, temperature and precipitation.

Figure 4 Data sets processing (see online version for colours)

As we study the link between weather conditions and demand variation, and the impact of including weather data in the demand forecasting model, we methodologically compare two nested prediction models. The first model contains only operations related covariates, while the second model contains both operations related variables and weather covariates from the previous year. There are a considerable number of published papers on demand forecasting. They are mainly based on time series analyses (Taylor, 2003), GLMs (Wasserman et al., 1991), neural networks (Al-Saba and El-Amin, 1999) or

machine learning techniques (Carbonneau et al., 2008). To analyse our data we opted for GAM, an extension of GLMs, since they let the data “speak for themselves” via nonparametric functions. The methods standardly used in operations management for demand forecasting are the linear regression or generalised linear regression models or moving average-based models such as the EWMA. We go one step further in terms of flexibility by letting the data decide for the functional form and use the GAMs that adequately capture in the same time seasonality, long term variation and other changes due to different levels of covariates. Therefore they can better handle non-linear relationships between the covariates and the response variable. This is essential when working with weather variables, since weather effects are not always linear. More precisely, GAMs are GLMs in which a transformation g of the expectation μ of the response variable linearly depends on unknown smooth functions of some covariates (Guisan et al., 2002). In other words, GAMs are semi-parametric extensions of GLMs (Hastie and Tibshirani, 1986, 1990) and the only underlying assumption made is that the functions are additive and the components are smooth. By applying GAMs, we model the mean weekly total order net value in a geographical area using the set of covariates x_i :

$$g(\mu) = \beta_0 + f_1(x_1) + f_2(x_2) + \dots + f_m(x_m)$$

where x_i ($i = 1, \dots, m$) denotes the covariates, β_0 the intercept, f_i smooth functions and g is the link function. The smooth functions may be specified with a parametric form, non-parametrically or semi-parametrically. This flexibility to allow non-parametric fits with relaxed assumptions on the actual relationship between a response variable and explanatory variables, provides the potential for better fits to data than parametric models, but arguably with some loss of quantitative interpretability although it allows a convenient visual result. Thus the strength of GAMs is in their ability to deal with highly non-linear and non-monotonic relationships between the response and the covariates. In our current case, the dependent variable is *order volume in CHF*, the possible covariates *previous year's order volume*, *order week*, *station*, and *previous year's weather variables*, namely, *snow*, *fresh snow*, *temperature*, *precipitation*, *humidity*, *pressure*, *sun* and *wind*. According to preliminary analyses, we assume that the order volume follows a gamma distribution (Burgin, 1975). Therefore, we decided to use the gamma family distribution (Dadpay et al., 2007) with the logarithm as the link function in our GAM models.

A crucial step in applying GAMs is to select the appropriate level of the smoother for the covariates. The smooth functions are flexible and could lead to an overfitting. A reasonable balance must be maintained between the total number of observations and the total number of degrees of freedom used when fitting the model. To find a suitable model, several nested models including different covariates were fitted and compared using the Akaike Information Criterion (AIC, see Akaike, 1973; Sakamoto et al. 1986):

$$AIC = -2(\ln(\text{likelihood})) + 2k,$$

where the likelihood is the probability of the data given a model and k is the number of free parameters in the model. AIC is based on information theory, where the underlying aim is to minimise the loss of information using Kullback-Leibler distance (Kullback and Leibler, 1951) between the model and the truth. Since the AIC is only valid asymptotically, we decided to use the AICc, which is a corrected AIC valid for finite

sample sizes (n) (Anderson and Burnham, 2002). The lowest AICc indicates the best model.

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

We tested several methods, such as principle component analysis (Jolliffe, 2002) and self-organising maps (Kohonen and Somervuo, 1998), to find the best set of candidates, and after that eliminating the least significant covariates one by one.

In this paper, for both countries we compare two nested prediction models, which are listed below. The first model contains only operations related covariates. The second model contains both operations related and previous year weather variables. The term ‘by’ in the final smooth function (f_{27}) indicates that for each level of the second covariate (for each level of the station) the model fits a smooth function to the main covariate (previous year daily sunshine duration).

1 Swiss prediction models:

- $M1$: $\log(\text{Order volume}) = \beta_{10} + f_{11}(\text{Previous year order volume}) + f_{12}(\text{Order week}) + f_{13}(\text{altitude})$
- $M2$: $\log(\text{Order volume}) = \beta_{20} + f_{21}(\text{Previous year order volume}) + f_{22}(\text{Order week}) + f_{23}(\text{altitude}) + f_{24}(\text{Previous year precipitation}_2 \text{ by weather season}) + f_{25}(\text{Previous year precipitation}_2 \text{ by station}) + f_{26}(\text{Previous year sun}_2 \text{ by weather season}) + f_{27}(\text{Previous year sun}_2 \text{ by station}) + f_{28}(\text{Previous year temperature}_{dv} \text{ by weather season}) + f_{29}(\text{Previous year temperature}_{dv} \text{ by station})$

2 Finnish prediction models:

- $FM1$: $\log(\text{Order volume}) = \beta_{10} + f_{11}(\text{Previous year order volume}) + f_{12}(\text{Order week}) + f_{13}(\text{Station})$
- $FM2$: $\log(\text{Order volume}) = \beta_{20} + f_{21}(\text{Previous year order volume}) + f_{22}(\text{Order week}) + f_{23}(\text{Station}) + f_{24}(\text{Previous year precipitation by weather season}) + f_{25}(\text{Previous year precipitation by station})$

4 Results

4.1 Model results

This section will present the results with an identifier (a) for Swiss results and (b) for Finnish ones. The results of the GAM models are summarised in Tables 3(a) and 3(b).

According to the AICc the Swiss model M2, which contains weather covariates, provides more information. The values for ‘deviance explained’, which is the percentage of the null deviance explained by the model, are 45 and 48.2%, for model M1 and M2 respectively. The analysis of variance using a chi squared test shows that the two models are significantly different (p -value $< 2e-16$). These results allow us to conclude that past weather conditions have an impact on seasonal demand. In other words, incorporating weather data, which in this case means considering precipitation, temperature and sun, leads to a better model. Moreover this means, if possible, one should integrate weather

available data in the demand forecasting process. The estimated degrees of freedom (EDF) and approximate significance levels (p -values) are displayed for each of the covariates in Table 4(a). Prior to obtaining Model 2, we tested several models mixing all weather variables available, including snow, fresh snow, temperature, precipitation, humidity, pressure, sun and wind. We found that humidity, snow, pressure and wind were not significant in the selected model. In the selected model, as temperature, we consider the deviation of the daily maximum to the long term average of the day. This covariate allows us to exclude the seasonal effect.

The AICc of the model Finnish model FM2 (83,239) is slightly lower than the AIC of FM1 (83,348), indicating that the model with weather covariates is better [Table 3(b)]. The deviance explained is higher for the second model (26.8 > 25.1), meaning that the second model explains a larger percentage of the null deviance. The analysis of variance using the chi squared test shows that the two models are significantly different by rejecting the null hypothesis with a p -value equal to 2.47e-06. As for the Swiss case, considering weather variables lead to a better model.

Table 3(a) Swiss GAM models' results

	<i>Model M1</i>	<i>Model M2</i>
N	7,488	7,488
AICc	141,475	141,091
Deviance explained (%)	45	48.2

Table 3(b) Finnish GAM models' results

	<i>Model FM1</i>	<i>Model FM2</i>
N	5,518	5,518
AICc	83,348	83,239
Deviance explained (%)	25.1	26.8

Table 4(a) Swiss GAM results: estimated degrees of freedom (EDF) and approximate significance levels (p -values)

<i>Covariates</i>	<i>Model M1</i>		<i>Model M2</i>	
	<i>EDF</i>	<i>p-value</i>	<i>EDF</i>	<i>p-value</i>
Previous year order volume	3.51	1.3e-07(***)	4.08	3.5e-12(***)
Order week	8.85	< 2e-16(***)	8.82	< 2e-16(***)
Altitude	8.96	< 2e-16(***)	8.93	< 2e-16(***)
Previous year precipitation_2 by weather season:	-	-		
• Autumn			2.44	0.097(.)
• Spring			3.45	0.015(*)

Notes: Significance codes for the p -value: 0 (***)0.001 (**).01 (*)0.05 (.)0.1

Table 4(a) Swiss GAM results: estimated degrees of freedom (EDF) and approximate significance levels (*p*-values) (continued)

<i>Covariates</i>	<i>Model M1</i>		<i>Model M2</i>	
	<i>EDF</i>	<i>p-value</i>	<i>EDF</i>	<i>p-value</i>
Previous year precipitation_2 by station:	-	-		
• BER			7.17	4.1e-16(***)
• CDF			4.67	0.0016(**)
• GVE			1.02	0.057(.)
• OTL			1.04	0.015(*)
• SAM			1.01	0.088(.)
• SMA			4.79	0.037(.)
• STG			3.77	0.023(*)
Previous year sun_2 by weather season:	-	-		
• Summer			5.61	0.027(*)
Previous year sun_2 by station:	-	-		
• CDF			2.16	0.072(.)
Previous year temperature_dv by weather season:	-	-		
• Spring			5.79	8.8e-05(***)
Previous year temperature_dv by station:	-	-		
• CDF			6.2	9.1e-08(***)

Notes: Significance codes for the *p*-value: 0 (***)0.001 (**)0.01 (*)0.05 (.)0.1

Table 4(b) Finnish GAM results: estimated degrees of freedom (EDF) and approximate significance levels (*p*-values)

<i>Covariates</i>	<i>Model FM1</i>		<i>Model FM2</i>	
	<i>EDF</i>	<i>p-value</i>	<i>EDF</i>	<i>p-value</i>
Previous year order volume	3.14	0.198	3.26	0.03(*)
Order week	8.74	<2e-16(***)	8.75	<2e-16(***)
Previous year precipitation by weather station:	-	-		
• 45100			5.7	0.09(.)
• 70100			1.0	0.04(*)
• 99600			7.8	2.09e-5(***)

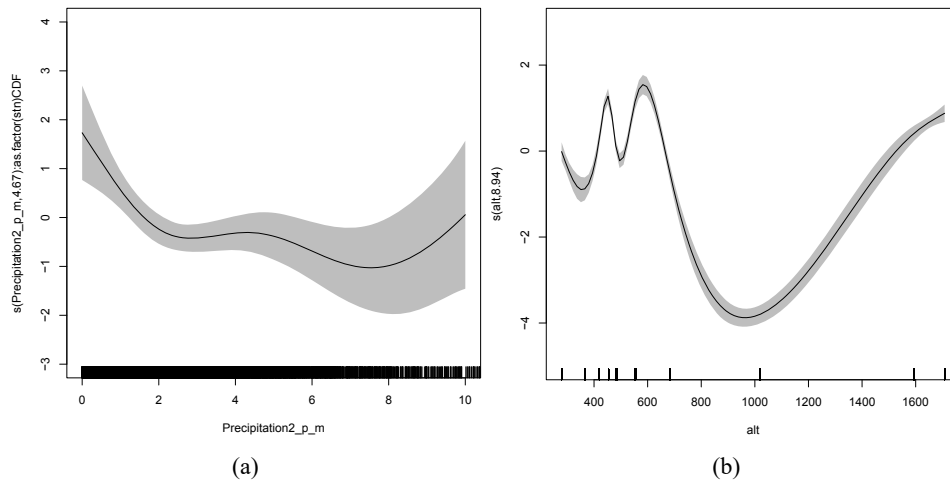
Notes: Significance codes for the *p*-value: 0 (***)0.001 (**)0.01 (*)0.05 (.)0.1

Next we study two covariates in more detail. In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls by gravity. The main form of precipitation includes drizzle, rain, sleet, snow and hail. The smooth function of the effect of daily sum of precipitation in La Chaux-de-Fonds is illustrated in Figure 5(a). Up to 2 mm, the effect of precipitation is positive but with a negative slope, meaning that as the

amount of precipitation grows its positive effects on order volume decrease. Above 2 mm, the effect of daily sum of precipitation on order volume is clearly negative, meaning that customers, including are retailers, sport shops etc., tend to order less when the amount of precipitation is higher than 2 mm.

The smooth function of altitude [Figure 5(b)] shows clearly the difference between urban areas and ski resorts in Switzerland. In urban areas, which correspond to 0–700 metres altitude, the effects of the altitude variable is undulating and oscillates around 0, meaning that altitude does not have a significant effect on orders. Unlike urban areas, in ski resorts (from 1000 m) the effect of altitude is almost linear with a positive slope. This means that customers in higher ski resorts tend to order more than those located in lower ones. Naturally the latter result matches the fact that altitude is positively correlated with snow.

Figure 5 Smooth functions of model M2 with 95% confidence intervals, illustrating the effects of, (a) precipitation in La Chaux-De-Fonds and (b) the effect of altitude (see online version for colours)

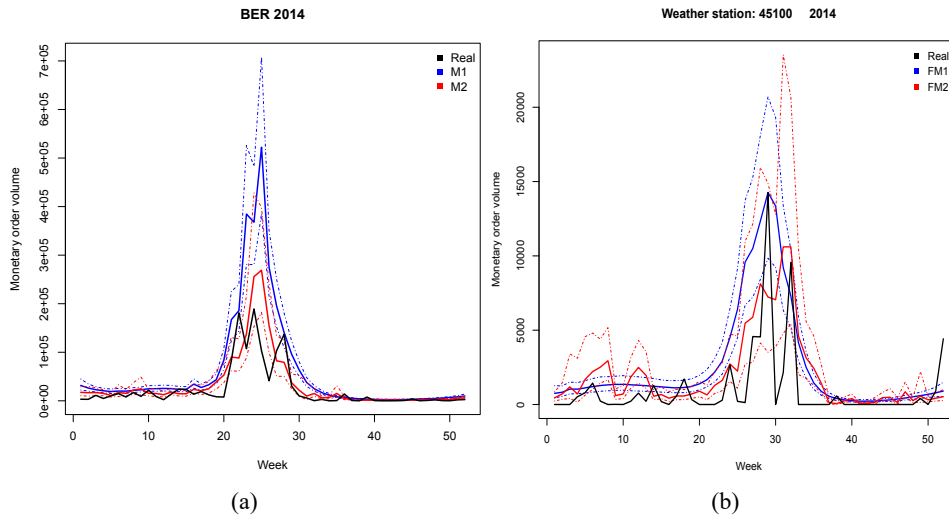


4.2 Prediction results

As mentioned previously, our Swiss dataset covers 12 years of operation data. For each year we use data from all previous years as a training set. For example, to predict the order volume for the 2013 sales season, we use seasons 2003–2012 to fit the models. The incremental improvement gained by including weather data in the forecasting process is clearly seen in 2013 and 2014. The percentage errors¹ (PE) decreases from 4.8% to 2.6% for 2013, and from 20.0% to 18.3% for 2014, this corresponds to a reduction in PE of –45% and –8.7%, respectively. For previous years, the models need more historical data to become stabilised. The values of weekly aggregated order volumes provided by different models for 2014 in Bern are illustrated in Figure 6(a) with 95% confidence intervals.

Figure 6(b) illustrates the values of weekly aggregated order volumes provided by different models for 2014 in the area corresponding to the Finnish weather station ‘45100’ with 95% confidence intervals.

Figure 6 (a) Weekly aggregated monetary volume of orders for 2014 in Bern, with 95% confidence intervals. comparison of the actual orders and the predicted orders from M1 and M2 models, the predictions provided by model M2 (red line) is closer to real order values (black line) (b) Weekly aggregated monetary volume of orders for 2014 in the area corresponding to weather station 45100, with 95% confidence intervals. Comparison of the actual orders and the predicted orders from FM1 and FM2 models, the predictions provided by model FM2 (red line) is closer to real order values (black line) (see online version for colours)



5 Discussion

The winter sports business is weather dependent to the degree that managers often raise their hands: “You can do as much analysis as you like, but in the end the snow comes or it doesn’t”. Reliable weather forecasts can extend up to ten days, while business decisions like production volumes need to be made up to 12 months in advance. However, our research demonstrates that smart managers do not need to run their operations blindly. A great winter is everyone’s hope in the alpine cluster: manufacturers, retailers, consumers, but also hotel owners and lift operators. There is a common understanding of what great winter weather is like: it is cold (but not too cold) and sunny with lots of snow. The first contribution of our research is the operationalisation of a ‘great winter’ with publicly available meteorological measurement data: temperature, precipitation, snow depth, and hours of sun. This work quantifies the somewhat fuzzy concept of nice skiing weather into something that can be measured and reported consistently.

As a second contribution, we have demonstrated that the weather in one winter affects sales in the next winter. This is a non-intuitive result, since traditionally, great winter weather has been expected to increase the in-season sales of ongoing winter, but not pre-season sales for the next winter. However, given the facts, the effect can be explained. A great winter means high sell-outs for retailers. They will need new inventory for the next season, rather than continuing to sell the last season’s leftovers. Great winters also inspire people to do more of their sport, raising expected consumer demand in the future. Finally, a great winter can have a psychological effect causing retailers to be more optimistic

when placing their orders for the next season, which follows Kahneman’s (2011) availability heuristic, that is, a mental shortcut that relies on immediate examples that come to a given person’s mind when evaluating a specific topic, concept, method or decision.

For supply chain management in the winter sports equipment business, the above two findings combined would support the management during the critical time window from January to April when major supply decisions for the next season are made. At that point in time, all orders and their deliveries for the ongoing season are known. However, these orders were placed in the previous season, which was about a year ago, for a different range of products and partly by a different set of customers. The orders for the next season are booked between January and April: typically depending on the schedule of sales visits. Hence, there is a time lag of up to three months until the weather of the current winter affects the order book for the next winter. This means that some of the orders for the next season are made before the current season has ended. Weather data, however, is available in real time and helps forecast order volumes. By using weather data, a winter sports company can improve the accuracy of their sales forecasts at the time when sales forecasts are most needed.

Figure 7 Time windows for different activities in the supply chain for seasonal winter sport products. Weather data follows seasonal sales period and customer order intake cycle

	Year -1					Year 0					Year +1												
	J	F	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N
Customer orders	Pre season 80%		Within season 20%			Pre season 80%		Within season 20%			Pre season 80%		Within season 20%										
Supplier orders	Customer order visibility from 3% to 60%					Customer order visibility from 3% to 60%					Customer order visibility from 3% to 60%												
Deliveries to customers	Peak volumes in Sept & Oct					Peak volumes in Sept & Oct					Peak volumes in Sept & Oct												
Sales period/ weather data	... season -1					Sales of season 0					Sales of season 1					... season 2							

In the modelled case of Switzerland, the forecasting accuracy improvement for full season sales value is up to two percentage points. Improving the forecasting accuracy by even one percentage point has a positive effect on customer service, gross margins, and inventories. This encourages the case company to continue their initiatives to become more data driven, and to use advanced analytics for driving critical business decisions.

6 Conclusions and future research

We studied how weather conditions impact demand of seasonal products and how weather data could be used to improve seasonal demand forecasting accuracy. We used aggregate weekly orders combined with customer location based weather conditions in Switzerland and Finland. We found that location based weather conditions have a significant impact on order volumes and that incorporating this data with the demand forecasting process increases the demand forecasting accuracy. The incremental improvement gained by including weather data in the forecasting process corresponds to a reduction in PE of -45% and -8% for 2013 and 2014, respectively. These results can be

used, for instance, to integrate the impact of variability in weather in the decision making process to better anticipate demand volumes and reduce costs due to excess inventory or stock shortages. In our case, in Switzerland, the value of forecasting error reductions is around 2% of total sales value in 2013. Therefore, the first contribution of our research is the operationalisation of a ‘great winter’ with publicly available meteorological measurement weather data such as: temperature, precipitation, snow depth, and hours of sun. The second contribution of this paper is the fact that we have demonstrated that the weather in one winter affects sales in the next winter. Since one usually expects a great winter weather to increase the in-season sales of ongoing winter, but not pre-season sales for the next winter, the results found are non-intuitive.

The theoretical contribution of this article is to extend the Newsvendor model to seasonal products when the lag between ordering and demand is one year and that a weather-adjusted model provides more accurate demand forecasts. In other words, these results contribute to the theory of stochastic inventory control by showing that taking into account external disturbances, in this case the fluctuations in weather, improves forecasting accuracy in situations where the lag between ordering and demand is around one year. As for the practical implications the paper shows that weather adjusted forecasts improve forecasting accuracy even for seasonal products. This in turn improves supply chain efficiency and customer satisfaction through reduced inventory costs and better punctuality in deliveries. The second contribution of this paper is related to the use of GAM models. In previous literature, the methods used were mainly GLM, machine learning and time series techniques (Aburto and Weber, 2007; Thomassey, 2010; Tabrizi and Ghaderi, 2016; Appelqvist et al., 2016). The strength of GAM models lies in their flexibility to allow non-parametric fits and the ability to handle highly nonlinear relationships via smooth functions.

The main limitation of this research is the fact that we study leisure goods, which can eventually limit the generalisation of the results. Even though, we analysed two countries, which are essentially different, it could be advantageous to consolidate the results by analysing additional countries in future research.

As mentioned earlier, our dataset concerns highly seasonal products, and we observed that this seasonality was mainly captured by the order week. For future research, we can go deeper in the analysis, and better handle this seasonality, we can consider ‘un-seasonalising’ the orders with time series techniques before applying the GAM models to the residuals. Additionally, in this paper, the weather data from previous years are treated equally in the training process. As an improvement, we could weight them by giving more importance to the most recent years. This is motivated by the fact that people tend to rely more heavily on the past couple of years when establishing their mental conception of how nice the year was in terms of weather. Other avenues for future research could include the exploration of how other external datasets could be used to improve demand forecasting accuracy for seasonal products.

References

- Aburto, L. and Weber, R. (2007) ‘Improved supply chain management based on hybrid demand forecasts’, *Applied Soft Computing*, Vol. 7, No. 1, pp.136–144.
- Akaike, H. (1973) ‘Maximum likelihood identification of Gaussian autoregressive moving average models’, *Biometrika*, Vol. 60, No. 2, pp.255–265.

- Al-Saba, T. and El-Amin, I. (1999) 'Artificial neural networks as applied to long-term demand forecasting', *Artificial Intelligence in Engineering*, Vol. 13, No. 2, pp.189–197.
- Amini, A. and Ghodsi, R. (2016) 'A linear mathematical model for a transportation-inventory problem in a two-stage supply chain with different types of fuels for vehicles', *International Journal of Services and Operations Management*, Vol. 25, No. 3, pp.347–360.
- Anderson, D.R. and Burnham, K.P. (2002) 'Avoiding pitfalls when using information-theoretic methods', *The Journal of Wildlife Management*, Vol. 66, No. 3, pp.912–918.
- Apergis, N., Gabrielsen, A. and Smales, L.A. (2016) '(Unusual) weather and stock returns – I am not in the mood for mood: further evidence from international markets', *Financial Markets and Portfolio Management*, Vol. 30, No. 1, pp.63–94.
- Appelqvist, P., Babongo, F., Chavez-Demoulin, V., Hameri, A.P. and Niemi, T. (2016) 'Weather and supply chain performance in sport goods distribution', *International Journal of Retail & Distribution Management*, Vol. 44, No. 2, pp.178–202.
- Aviv, Y. (2001) 'The effect of collaborative forecasting on supply chain performance', *Management Science*, Vol. 47, No. 10, pp.1326–1343.
- Bahng, Y. and Kincade, D.H. (2012) 'The relationship between temperature and sales: sales data analysis of a retailer of branded women's business wear', *International Journal of Retail & Distribution Management*, Vol. 40, No. 6, pp.410–426.
- Behe, B.K., Getter, K.L. and Yue, C. (2012) 'Should you blame the weather? The influence of weather parameters, month, and day of the week on spring herbaceous plant sales in the US Midwest', *HortScience*, Vol. 47, No. 1, pp.71–73.
- Bertrand, J.L., Brusset, X. and Fortin, M. (2015) 'Assessing and hedging the cost of unseasonal weather: case of the apparel sector', *European Journal of Operational Research*, Vol. 244, No. 1, pp.261–276.
- Brockett, P.L., Wang, M. and Yang, C. (2005) 'Weather derivatives and weather risk management', *Risk Management and Insurance Review*, Vol. 8, No. 1, pp.127–140.
- Burgin, T.A. (1975) 'The gamma distribution and inventory control', *Journal of the Operational Research Society*, Vol. 26, No. 3, pp.507–525.
- Busse, M.R., Pope, D.G., Pope, J.C. and Silva-Risso, J. (2015) 'The psychological effect of weather on car purchases', *The Quarterly Journal of Economics*, Vol. 130, No. 1, pp.371–414.
- Carbonneau, R., Laframboise, K. and Vahidov, R. (2008) 'Application of machine learning techniques for supply chain demand forecasting', *European Journal of Operational Research*, Vol. 184, No. 3, pp.1140–1154.
- Chen, F.Y. and Yano, C.A. (2010) 'Improving supply chain performance and managing risk under weather-related demand uncertainty', *Management Science*, Vol. 56, No. 8, pp.1380–1397.
- Chopra, S. and Lariviere, M.A. (2005) 'Using service inventory to push performance', *Sloan Management Review*, Vol. 47, No. 1, pp.56–63.
- Christopher, M., Lowson, R. and Peck, H. (2004) 'Creating agile supply chains in the fashion industry', *International Journal of Retail & Distribution Management*, Vol. 32, No. 8, pp.367–376.
- Costantino, F., Di Gravio, G., Shaban, A. and Tronci, M. (2013) 'Exploring the bullwhip effect and inventory stability in a seasonal supply chain', *International Journal of Engineering Business Management*, Vol. 5, No. 23, pp.1–12.
- Cui, K. and Swishchuk, A. (2015) 'Applications of weather derivatives in the energy market', *The Journal of Energy Markets*, Vol. 8, No. 1, p.59.
- Dadpay, A., Soofi, E. S. and Soyer, R. (2007) 'Information measures for generalized gamma family', *Journal of Econometrics*, Vol. 138, No. 2, pp.568–585.
- Devinney, T.M. and Stewart, D.W. (1988) 'Rethinking the product portfolio: a generalized investment model', *Management Science*, Vol. 34, No. 9, pp.1080–1095.

- Fisher, M., Hammond, J., Obermeyer, W. and Raman, A. (1997) 'Configuring a supply chain to reduce the cost of demand uncertainty', *Production and Operations Management*, Vol. 6, No. 3, pp.211–225.
- Forrester, J. (1961) *Industrial Dynamics*, MIT Press/Wiley, New York.
- Gerchak, Y. (2016) 'Manufacturing newsvendors and inventory pooling with nonlinear production costs', *International Journal of Inventory Research*, Vol. 3, No. 1, pp.70–80.
- Goetzmann, W.N., Kim, D., Kumar, A. and Wang, Q. (2014) 'Weather-induced mood, institutional investors, and stock returns', *Review of Financial Studies*, hhu063.
- Guisan, A., Edwards, T.C. and Hastie, T. (2002) 'Generalized linear and generalized additive models in studies of species distributions: setting the scene', *Ecological Modelling*, Vol. 157, No. 2, pp.89–100.
- Hastie, T. and Tibshirani, R. (1986) 'Generalized additive models', *Statistical Science*, Vol. 1, No. 3, pp.297–310.
- Hastie, T.J. and Tibshirani, R.J. (1990) *Generalized Additive Models*, Vol. 43, CRC Press, Boca Raton.
- Huurman, C., Ravazzolo, F. and Zhou, C. (2012) 'The power of weather', *Computational Statistics & Data Analysis*, Vol. 56, No. 11, pp.3793–3807.
- Ishikawa, A. and Nejo, T. (1998) *The Success of 7-Eleven Japan: Discovering the Secrets of the World's Best-run Convenience Chain Stores*, World Scientific Publishing Co., Singapore.
- Jolliffe, I. (2002) *Principal Component Analysis*, John Wiley Sons, Ltd., New York.
- Kahneman, D. (2011) *Thinking Fast and Slow*, Farrar, Strauss, Giroux, New York, NY.
- Kleiderman, A. (2004) *Soggy Summer Spells Boardroom Gloom*, 24 September, BBC News.
- Kohonen, T. and Somervuo, P. (1998) 'Self-organizing maps of symbol strings', *Neurocomputing*, Vol. 21, No. 1, pp.19–30.
- Kullback, S. and Leibler, R.A. (1951) 'On information and sufficiency', *The Annals of Mathematical Statistics*, Vol. 22, No. 1, pp.79–86.
- Lee, H.L. and Tang, C.S. (1997) 'Modelling the costs and benefits of delayed product differentiation', *Management Science*, Vol. 43, No. 1, pp.40–53.
- Lee, H.L., Padmanabhan, V. and Whang, S. (2004) 'Information distortion in a supply chain: the bullwhip effect', *Management Science*, Vol. 50, No. 12, Supplement, pp.1875–1886.
- Leon, S., Szmerekovsky, J. and Tolliver, D. (2015) 'Using VAR for strategic capacity allocation: an airline perspective', *International Journal of Services and Operations Management*, Vol. 21, No. 2, pp.127–149.
- Levy, O. and Galili, I. (2008) 'Stock purchase and the weather: individual differences', *Journal of Economic Behavior & Organization*, Vol. 67, No. 3, pp.755–767.
- Lu, J. and Chou, R.K. (2012) 'Does the weather have impacts on returns and trading activities in order-driven stock markets? Evidence from China', *Journal of Empirical Finance*, Vol. 19, No. 1, pp.79–93.
- Ma, Y., Wang, N., Che, A., Huang, Y. and Xu, J. (2013) 'The bullwhip effect on product orders and inventory: a perspective of demand forecasting techniques', *International Journal of Production Research*, Vol. 51, No. 1, pp.281–302.
- Murray, K.B., Di Muro, F., Finn, A. and Leszczyc, P.P. (2010) 'The effect of weather on consumer spending', *Journal of Retailing and Consumer Services*, Vol. 17, No. 6, pp.512–520.
- Nicola, F. (2015) 'The impact of weather insurance on consumption, investment, and welfare', *Quantitative Economics*, Vol. 6, No. 3, pp.637–661.
- Porteus, E.L. (1990) 'Stochastic inventory theory', in Heyman, D.P. and Sobel, M.J. (Eds.): *Handbooks in Operations Research and Management Science*, Vol. 2, pp.605–652, Elsevier, Amsterdam.

- Sakamoto, Y., Ishiguro, M. and Kitagawa, G. (1986) 'Akaike information criterion statistics', *Dordrecht*, D. Reidel, The Netherlands.
- Sevlian, R. and Rajagopal, R. (2014) *Short Term Electricity Load Forecasting on Varying Levels of Aggregation*, Working Paper, arXiv preprint arXiv:1404.0058.
- Singh, R.K. and Acharya, P. (2015) 'Reverse supply chain flexibility: a theoretical framework of research dimensions', *International Journal of Services and Operations Management*, Vol. 22, No. 4, pp.442–454.
- Starr-McCluer, M. (2000) *The Effects of Weather on Retail Sales*, Federal Reserve Board of Governors, Washington D.C. [online] <http://www.federalreserve.gov/pubs/feds/2000/200008/200008pap.pdf>.
- Steinker, S. and Hoberg, K. (2015) *The Value of Weather Information for E-commerce Operations*, Working paper, Kühne Logistics University – KLU, Hamburg, Germany.
- Sting, F.J. and Huchzermeier, A. (2014) 'Operational hedging and diversification under correlated supply and demand uncertainty', *Production and Operations Management*, Vol. 23, No. 7, pp.1212–1226.
- Tabrizi, B.H. and Ghaderi, S.F. (2016) 'Sales forecasting of a dairy product manufacturing company: a comparative study of autoregressive integrated moving average and local linear neuro-fuzzy models', *International Journal of Services and Operations Management*, Vol. 24, No. 4, pp.531–547.
- Taylor, J.W. (2003) 'Short-term electricity demand forecasting using double seasonal exponential smoothing', *Journal of the Operational Research Society*, Vol. 54, No. 8, pp.799–805.
- Taylor, T.A. and Xiao, W. (2010) 'Does a manufacturer benefit from selling to a better-forecasting retailer?', *Management Science*, Vol. 56, No. 9, pp.1584–1598.
- Thomas, H.R., Riley, D.R. and Sanvido, V.E. (1999) 'Loss of labor productivity due to delivery methods and weather', *Journal of Construction Engineering and Management*, Vol. 125, No. 1, pp.39–46.
- Thomassey, S. (2010) 'Sales forecasts in clothing industry: the key success factor of the supply chain management', *International Journal of Production Economics*, Vol. 128, No. 2, pp.470–483.
- Treanor, S.D., Simkins, B.J., Rogers, D.A. and Carter, D.A. (2014) 'Does operational and financial hedging reduce exposure? Evidence from the US airline industry', *Financial Review*, Vol. 49, No. 1, pp.149–172.
- Van der Vorst, J.G. and Beulens, A.J. (2002) 'Identifying sources of uncertainty to generate supply chain redesign strategies', *International Journal of Physical Distribution & Logistics Management*, Vol. 32, No. 6, pp.409–430.
- Van der Vorst, J.G.A.J., Beulens, A.J., Wit, W.D. and Beek, P.V. (1998) 'Supply chain management in food chains: improving performance by reducing uncertainty', *International Transactions in Operational Research*, Vol. 5, No. 6, pp.487–499.
- Van Mieghem, J.A. (2007) 'Risk mitigation in newsvendor networks: resource diversification, flexibility, sharing, and hedging', *Management Science*, Vol. 53, No. 8, pp.1269–1288.
- Wagner, S.M. and Bode, C. (2006) 'An empirical investigation into supply chain vulnerability', *Journal of Purchasing and Supply Management*, Vol. 12, No. 6, pp.301–312.
- Wasserman, J., Manning, W.G., Newhouse, J.P. and Winkler, J.D. (1991) 'The effects of excise taxes and regulations on cigarette smoking', *Journal of Health Economics*, Vol. 10, No. 1, pp.43–64.

Notes

1 $PE = abs\left(\frac{F - A}{A}\right)$

where

A_i actual value

F_i forecast value

Chapter 4

Forecasting (un-)seasonal demand
using geostatistics,
socio-economic and weather data

Forecasting (un-)seasonal demand using geostatistics, socio-economic and weather data

Flora Babongo*, Tapio Niemi,
Valérie Chavez-Demoulin and
Ari-Pekka Hameri

Faculty of Business and Economics,
University of Lausanne,
Lausanne, Switzerland
Email: Flora.BabongoBosombo@unil.ch
Email: tapio.niemi@unil.ch
Email: Valerie.Chavez@unil.ch
Email: Ari-Pekka.Hameri@unil.ch
*Corresponding author

Patrik Appelqvist

Faculty of Management,
Aalto University,
P.O. Box 11000, FI-00076 Aalto, Finland
Email: Patrik.Appelqvist@unisport.com

Abstract: Accurate demand forecasts are essential to supply chain management. We study the spatial demand variation of seasonal and unseasonal sport goods and demonstrate how demand forecast accuracy can be improved by using geostatistics and linking socio-economic and weather data with order line specific supply chain transactions. We found that the socio-economic features impact the demand of both seasonal and unseasonal products and unseasonal products are impacted more. Weather conditions affect only seasonal products. Cross-validation analyses show that using external information improves demand forecasting accuracy by reducing forecasting error up to 48%. The results can be applied both to the operational demand planning process and to the strategy used when making location-based decisions on supply chain actions, for example, deciding locations for new stores or running marketing campaigns.

Keywords: demand forecasting; seasonal products; socio-economic features; weather; geostatistics; kriging; semivariogram.

Reference to this paper should be made as follows: Babongo, F., Niemi, T., Chavez-Demoulin, V., Hameri, A-P. and Appelqvist, P. (2019) 'Forecasting (un-)seasonal demand using geostatistics, socio-economic and weather data', *Int. J. Business Forecasting and Marketing Intelligence*, Vol. 5, No. 1, pp.103–124.

Biographical notes: Flora Babongo is a PhD student at the University of Lausanne in Faculty of Business and Economics. She received her Bachelor's in Economics from the University of Lausanne and Master's in Statistics from

University of Neuchâtel. Before starting her PhD, she worked two years for the Swiss Federal Office of Statistics. She is interested in applying statistics in different fields.

Tapio Niemi holds a PhD in Computer Science from the University of Tampere, in 2001. He worked at the University of Tampere as a Scientist (1997–2001) and as an Assistant Professor (2002–2004). Since 2004, he has worked in the Technology Programme of the Helsinki Institute of Physics (HIP) at CERN, and since 2012 also as a Senior Scientist in Department of Operations at HEC Lausanne, Switzerland, focusing on green computing, data analytics and business intelligence.

Valérie Chavez-Demoulin is a Professor of Statistics at the HEC Lausanne, specialising in statistical methods for quantitative risk management in general, and the statistical modelling of extreme events in particular. More recent methodological work concerns conditional dependence structures modelling, non-parametric Bayesian models, dynamic extreme value theory models and extremes for non-stationary time series. Following her PhD in Statistics at the EPFL, she obtained a grant for a postdoctoral position in collaboration with the SLF in Davos. Afterwards, she has been a Research Fellow at the Department of Mathematics at ETH, Zurich. Aside from her research, she has been the quantitative risk manager for the Hedge Fund for three years. She is a member of the RiskLab, ETH, Zurich and an elected member of The International Statistical Institute (ISI).

Ari-Pekka Hameri is a Full Professor of Operations Management at the University of Lausanne, Switzerland. He has been involved with numerous international research and consulting projects dealing with production and supply chain management. He has published over 70 articles in international management journals concerning the management of production, projects and supply chains.

Patrik Appelqvist was responsible for supply chain and operations development at the Amer Sports, one of the leading sporting goods companies in the world. He is currently collaborating with the Department of Management of Aalto University.

This paper is a revised and expanded version of a paper entitled ‘Geospatial analysis of seasonal and unseasonal demand’ presented at International Society for Business and Industrial Statistics Conference; New York, USA, 7–9 June 2017.

1 Introduction

All manufacturers and retailers wish to have the most accurate demand forecasting. This paper aims to improve the demand forecasting of seasonal and unseasonal products using the fundamental concepts of spatial dependence and interpolation; and the incorporation of the socio-economic aspects and weather conditions impacts in the spatial dependence structure.

We focus on studying the demand fluctuations of seasonal and unseasonal leisure goods and on improving the accuracy of demand forecasts by integrating the spatial

dimension in the planning process. As seasonal products, we consider winter sport goods and as unseasonal products we consider indoor sports and golf equipment.

We analyse the real demand data of different products at different levels of aggregation, namely, at the brand and product family levels. At the brand level, the studied brands are Atomic and Wilson, the latter being unseasonal and the former seasonal. At the product family level, we analyse two non-overlapping product families for both brands mentioned above. The studied products are Atomic Alpine skis, Atomic X-Country skis, Wilson Racquet Sports equipment and Wilson Golf equipment. The two main aims of this research are to determine:

- 1 How demand varies geographically according to socio-economic aspects and weather conditions.
- 2 How the additional information, external to supply chain itself, affects demand forecasting accuracy.

To do this, we use order line data on several products of well-known brands owned by a listed company delivering sporting equipment and goods worldwide. The company dataset covers about 12 years of orders, which corresponds to more than 890,000 order lines. The spatial interpolation is performed with socio-economic data such as the number of inhabitants, the number of hotel nights, the size of households on average and the number of jobs along with weather conditions (snow, temperature, precipitation and sunshine).

The main purpose of our study is to explain, using model-based geostatistics (Diggle et al., 2003), how considered external information, namely socio-economic characteristics and weather conditions affect the spatial variation of demand. Our detailed hypotheses are:

- 1 Product specific demand is spatially impacted by socio-economic features.
- 2 Demand for seasonal products is spatially affected by weather conditions.
- 3 Incorporating socio-economic aspects and weather conditions in the demand prediction process improves the demand forecast accuracy.

As the main results of the study, we show that:

- 1 The spatial correlation of the demand for seasonal products, that is, the correlation of seasonal demand between different locations, is impacted by weather conditions.
- 2 The accuracy of demand forecasting can be improved if the weather information and socio-economic features are included in the model.

The results can be useful in the decision making, such as planning future demand to optimise inventories and orders, or deciding on the location of a new retail shop.

The rest of the paper is structured as follows. We start with a literature review on geostatistics applications in general and especially those used in demand forecasting in Section 2. We then formulate, in Section 3, our detailed research questions, provide the description of the datasets, followed by a descriptive analysis of the variables and the explanation of the applied statistical methodology used in the analysis. The results are provided, explained and discussed in Section 4. Finally, conclusions are drawn and avenues for future research are presented in Section 5.

2 Literature review

2.1 Socio-economic environment and weather conditions in demand forecasting

Literature on demand forecasting is wide thus we can only focus on a sample of studies most relevant to our topic. Polebitski and Palmer (2009) develop regression-based water demand models capable of forecasting single-family residential water demands using demographic, economic and weather data. Subsequently, Gage and Cooper (2015) assess the relative importance of physical and socio-economic variables in predicting outdoor water usage. They provide analyses and comparisons of the predictive accuracy of developed models. The models study different subsets of explanatory variables (see also Jain and Ormsbee, 2002).

Fadiga et al. (2005) identify sources of demand growth by analysing consumer demographic profiles, regions and product characteristics. They analyse nine apparel products through detailed demographic and socio-economic factors. The socio-economic data has been less explored spatially with regards to improving demand forecasting accuracy.

Regnier (2008) describes what advances in weather forecasting can offer and how weather information can be applied to operations research models for improving decision making. Additionally, sophisticated firms such as Fedex, UPS and various agriculture and energy companies more commonly employ meteorologists to improve their ability to forecast and to use those forecasts in making decisions (Lustgarten, 2005).

2.2 Geostatistics applications in various fields

What we now regard as geostatistics models and techniques were largely developed by Matheron (1963) to evaluate recoverable reserves for the mining industry. Therefore, it is not surprising to observe that geostatistics are mainly applied in fields that are directly related to soil, for instance the mining industry (Benndorf, 2014), soil science, geology, the coal industry (Srivastava, 2013), etc.

Soil has been widely studied using geostatistics techniques. McBratney (1992) analyses soil variation which is usually considered as problematic to optimal soil management. Geostatistical methods are used to investigate the spatial characteristics of the compaction data from several projects, with the goal of using these techniques to guide the quality assurance process (Petersen et al., 2007). Caeiro et al. (2003) use a set of multivariate geostatistical approaches to delineate spatially contiguous regions of sediment structure and Moral et al. (2010) characterise the spatial variability of the main soil physical variables.

In the field of agriculture, Oliver (2010) describes the two core techniques of geostatistics and illustrates their applications. *Variography* evaluates the evolution of the spatial correlation between two points when their separation distance increases. *Kriging* is an interpolation method (Kriging, 1951). The role of field-scale experiments in location-specific crop management is studied by Pringle et al. (2004a, 2004b) using different forms of kriging and the outcome is a map for each field that describes the optimum application of experimental input. Location-based crop management has widely benefited geostatistics (Inamura et al., 2004; Moral et al., 2011). Morari et al. (2009) analyse the spatial variability of soil properties and their relationships with electrical

conductivity in horizontal and vertical mode is estimated using multivariate geostatistical techniques.

Zhang et al. (2016) and Legleiter and Kyriakidis (2008) study the spatial interpolation of river channel topography using the shortest temporal distance, inverse distance weighting and kriging. Warner et al. (2006) discuss a stochastic groundwater model for the management of a pump and treat system located in the Offpost Operable Unit at the Rocky Mountain Arsenal.

Many other fields have benefited from geostatistic methods. Tuominen et al. (2003) combine the k -nearest-neighbours estimation, stand inventory data and geostatistical interpolation for the estimation of five forest variables (mean diameter, mean height, mean age, basal area and volume) per sample plot and stand. Nelson et al. (1999) focus on the spatial relationships of landscape features that interact with the progression of an epidemic to refine cultural management strategies for plant disease control. To evaluate the tsetse fly that poses a major constraint to crop and livestock production, Sciarretta et al. (2005) use geostatistical methods to identify and monitor the spatiotemporal dynamics of areas with increased fly densities, considered as hot spots. Namysłowska-Wilczyńska and Wilczyński (2015) analyse the superficial variability of electric power using lognormal and ordinary kriging that allows the identification of their nature and their range. Stelzenmüller et al. (2008) examine the spatial dynamic of artisanal fishing fleets around five European marine protected areas with geostatistical modelling techniques. They find that in most cases the factors ‘distance to the no-take’, ‘water depth’ and ‘distance to the port’ have a significant influence on effort allocation by the fishing fleets. In risk assessment and decision making, Rendu (2002) uses geostatistical simulation in the evaluation of the geologic risk in order to contribute in the assessment of the expected utility of a project.

2.3 Geostatistics applied to demand forecasting

There are many applications of geostatistical techniques in demand forecasting. Gomes et al. (2016) compare the results of two techniques, namely ordinary and indicator kriging, in the estimation of private motorised travel (car or motorcycle) in several geographical locations. They conclude that spatial statistics are thriving in travel demand forecasting issues. In public transportation, Prasetyowati et al. (2016) provide an analytical tool such as a model that can be used to govern a public policy regarding traffic management and help to solve issues such as how to determine public transportation routes, how to determine the type of public transportation and how to determine the optimal amount of public transportation needed for each route. They use the ordinary kriging to predict the occupancy of public transportation systems for each crowd spots area.

In water resource management, Muthuwatta et al. (2010) conduct a study to assess water availability and consumption in the Karkheh River Basin. They estimate the precipitation using geostatistical techniques, while a surface energy balance approach is selected for evapotranspiration estimation. Their results suggest that the water balance is sufficiently understood for policy and decision making.

Geostatistical methods are applied to forecasting electrical power demand (Namysłowska-Wilczyńska and Wilczyński, 2010). These authors analyse two kinds of electrical power networks using the directional variogram function and power demand

forecasts for one year and five years are made using ordinary block kriging. Their results show that the employed techniques are useful and effective.

In agribusiness, the market prices of crops are indicators for their demand and supply. Peng et al. (2015) investigate the spatial relationships between prices in different markets. To do so, they examine the forecasting price given by four well-known spatial algorithms, which are the nearest neighbour, inverse distance weighting, the kriging method and artificial neural networks. They finally compare the performance of these four algorithms with the price data obtained from 15 markets on the official website of the Council of Agriculture of Taiwan. They find that kriging leads to the lowest error in percentages. Considering the time efficiency, the kriging method is also recommended for the development of forecasting service, since the regression can be accomplished efficiently.

According to this literature review, geostatistics is an interesting tool that has been applied to various fields but not directly to supply chain demand forecasting, although geostatistics could offer clear benefits to supply chain management and demand forecasting. Therefore, in this study, we focus on applying geostatistical techniques to supply chain management by analysing the demand of seasonal and unseasonal products, incorporating socio-economic data and weather conditions in the modelling process.

3 Research questions, data and methodology

3.1 Research questions

One of the main questions for a retail company is how to choose the right site for a store, since the location determines the external market conditions. Our study analyses how using information on these external conditions can improve demand forecasting accuracy when applying geostatistical models. By analysing real order data on several, seasonal and unseasonal sport products, we show how external data explains spatial variations in demand and whether it is possible to make accurate forecasts, for example, for a new store location. The main research questions of the study are:

- How does demand geographically vary according to socio-economic aspects and weather conditions?
- Do external effects, namely socio-economic aspects and weather conditions, have a different impact on seasonal and unseasonal products demand?
- How does additional information, external to the supply chain, affect demand forecasting accuracy?

3.2 Data

The analysed data is from the Amer Sports Corporation, which is one of the leading sporting goods companies in the world with a sales network covering 34 countries and net sales over 2.6 billion euros. Customers of Amer Sports are mainly sporting goods chains, specialty retailers, mass merchants, fitness clubs and distributors. In this study, we focus on customers located in Switzerland. The company owns several global brand business units and provides its customers with a wide range of sports equipment for

indoor and outdoor sports for winter and summer. In this paper, we consider two different brands: Atomic as seasonal and Wilson as unseasonal.

- Atomic: The Atomic brand concerns skiing equipment, including both alpine and cross-country skis, ski boots, bindings, helmets, ski poles and goggles. At the product level, we analyse Alpine and X-Country skis.
- Wilson: This brand is involved in ball sports such as tennis, baseball, American football, golf, basketball, softball, badminton and squash. The company is structured into three business areas: racquet sports, team sports and golf. At the product level, we analyse racquet sports and golf equipment.

The corresponding dataset contains more than 890,000 orders made by more than 1,700 customers between 2003 and 2014. By analysing the temporal dimension, meaning the effect of the year on the spatial trend, we found that it is not significant. This means that after removing the trend, the studied spatial structure should not significantly differ from one year to another. To obtain this result, we remove the fitted spatial trend and the residuals analysis for different years. Using a chi-squared test shows that we cannot reject the null hypothesis, which corresponds to the fact that the sets of residuals from different years are not significantly different from each other ($p\text{-value} = 1$). This result indicates that the effect of the year on the spatial trend is not significant. We therefore focus on the analysis of the most recent data.

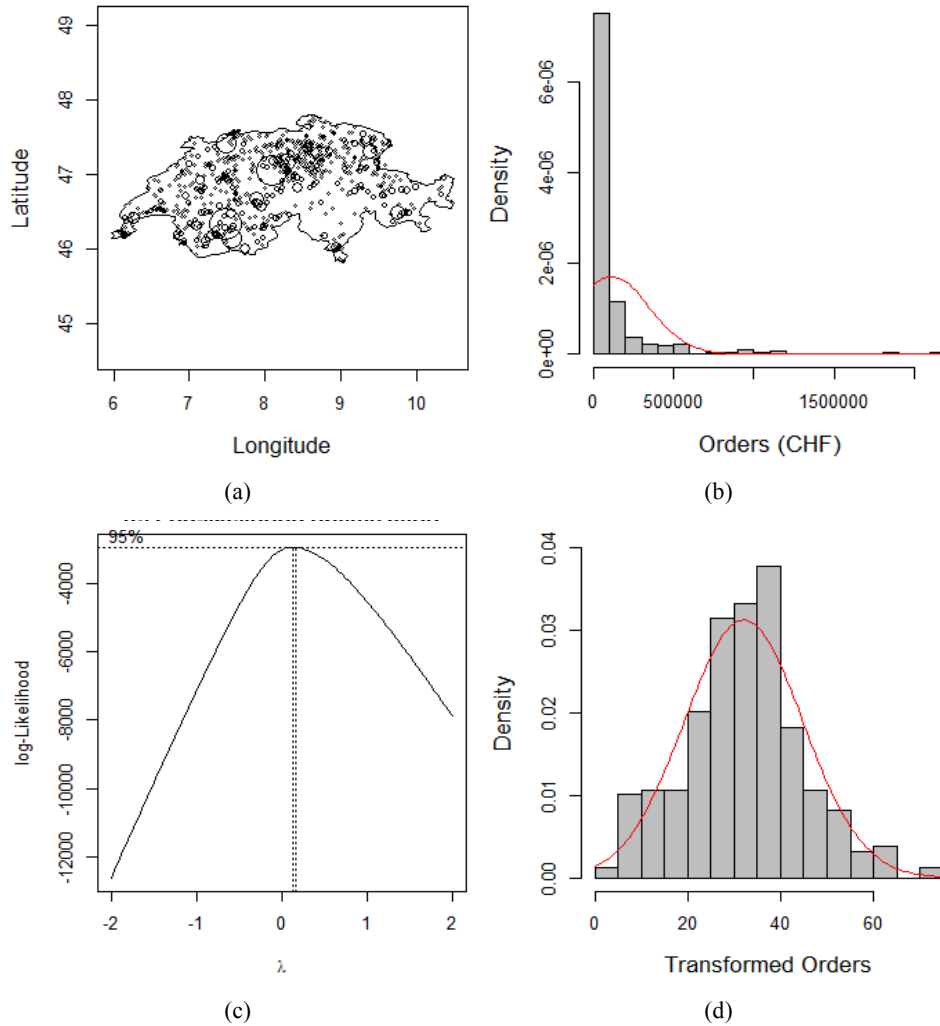
In the next step, this data, corresponding to the year 2014, is geographically matched with socio-economic data, retrieved from the Swiss Federal Statistical Office and weather data from the Swiss national weather database. The socio-economic dataset contains the following variables: number of inhabitants, population density per km^2 , percentage of population between 20–64 years old, the number of hotel nights, number of private households, percentage of housing environment and infrastructure surface, social assistance rate, number of companies in the primary, secondary and tertiary sectors, number of jobs in the primary, secondary and tertiary sectors. The weather dataset contains the following variables: gust, snow, pressure, temperature, precipitation, humidity, sun and wind.

The final dataset was obtained by merging the company, weather and socio-economic datasets according to the postal code. The further coming model selection provides the listed significant variables (Table 1).

Table 1 List of used variables

Order variable	Socio-economic variables per postal code
<ul style="list-style-type: none"> • Order volume value: in Swiss francs (CHF) 	<ul style="list-style-type: none"> • Number of hotel nights
Location coordinates: (long, lat.)	<ul style="list-style-type: none"> • Percentage of population between 20-64 years old (%)
<ul style="list-style-type: none"> • Longitude (°) • Latitude (°) • Altitude (m) 	<ul style="list-style-type: none"> • Private households • Percentage of housing environment and infrastructure surface (%)
Weather conditions variables (annual maximum)	<ul style="list-style-type: none"> • Number of companies in the primary sector
<ul style="list-style-type: none"> • Snow: thickness of snow measured at 05:40 am (cm) 	

Figure 1 (a) Atomic order value according to location (b) Histogram of the positively skewed distribution of Atomic orders with a normal fitted curve (c) Log-likelihood function with the chosen λ (d) Histogram of the distribution of transformed orders with a normal fitted curve (see online version for colours)



Note: (a) Order volume, (b) order volume distribution, (c) log-likelihood optimisation and (d) transformed order volume distribution.

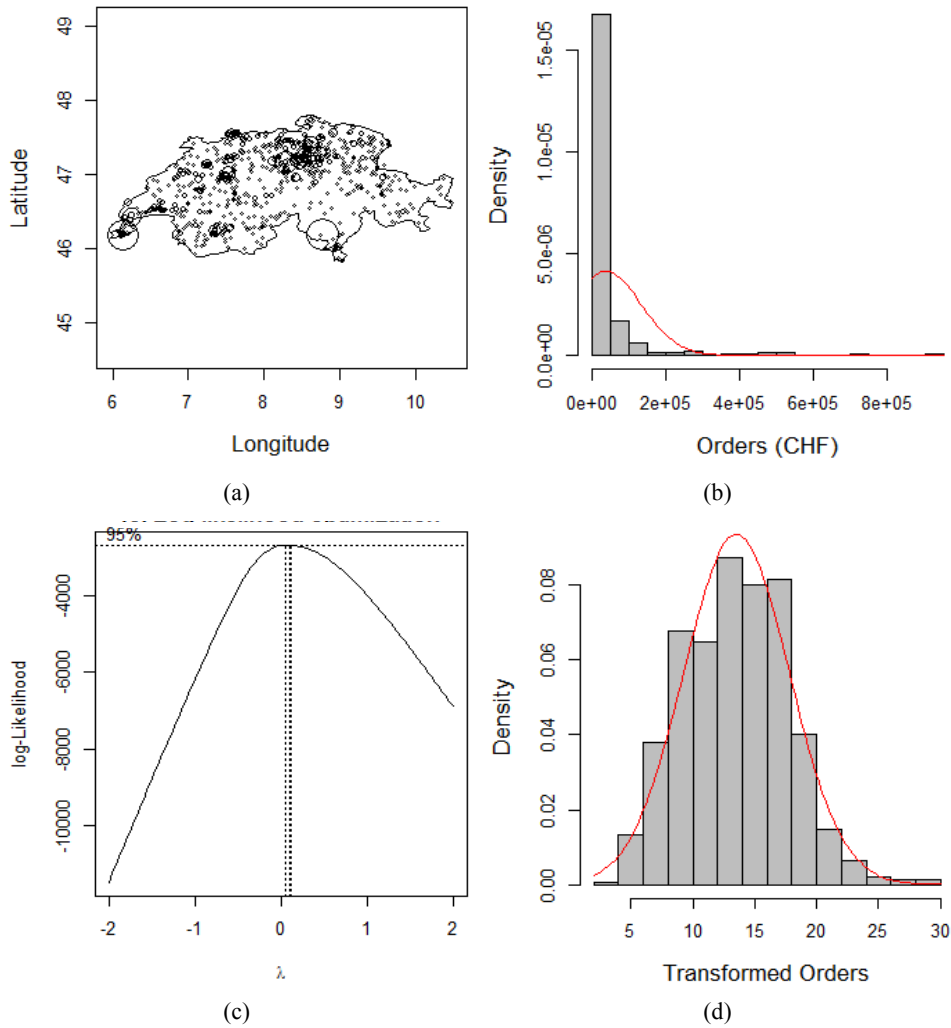
The response variable expressing demand is the order value measured in Swiss francs. We start with the univariate data analysis of the orders and exclude a maximum of two outliers depending on the dataset (Atomic, Atomic Alpine, Atomic X-Country, Wilson, Wilson Racketsports or Wilson Golf). The distribution of the Atomic (seasonal) and Wilson (unseasonal) orders values in Swiss francs (CHF) according to the location is displayed in Figures 1(a) and 2(a) along with the distribution of orders [Figures 1(b) and 2(b)]. The order distribution is positively skewed. To meet the relevant theoretical assumptions relating to the geostatistical model, namely the normality and the homoscedasticity (random variables having equal variances) of the data, we use the

Box-Cox transformation (Box and Cox, 1964), which reduces anomalies such as non-additivity, non-normality and heteroscedasticity. The transformed orders are calculated as follows:

$$Transformed\ orders = \begin{cases} \frac{Orders^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(Orders), & \lambda = 0 \end{cases}$$

The parameter λ is chosen such that it maximises the log-likelihood [Figures 1(c) and 2(c)]. Figures 1(d) and 2(d) illustrate the distribution of the transformed Atomic and Wilson orders with the normal fitted curve.

Figure 2 (a) Wilson order value according to location (b) Histogram of the positively skewed distribution of Wilson orders with a normal fitted curve (c) Log-likelihood function with the chosen λ (d) Histogram of the distribution of transformed orders with a normal fitted curve (see online version for colours)



3.3 Methodology

The main focus of the geostatistical data analysis is understanding and describing the spatial patterns of the response variable. The geostatistical data consists of the response Y (transformed orders in our case) associated with locations x (defined by the longitude and the latitude) in a continuous spatial region A (Switzerland).

We start with a short spatial exploratory analysis to empirically describe the correlation structure. A common tool used to describe this spatial dependence in geostatistics for exploratory purposes is the semivariogram, which describes the spatial association as a function of the separation distance.

The empirical semivariogram $\gamma(u)$ is given by equation below, where u is the distance interval and N_u is the total number of sample pairs within the distance interval u . The semivariogram shows the degradation of the spatial correlation between two points of space when the separation distance increases.

$$\gamma(u) = \frac{1}{2N_u} \sum_{i=1}^{N_u} [Y(x_{i+u}) - Y(x_i)]^2$$

We then continue with the model selection based on the main assumption which is that each observed value Y is either a direct measurement of, or is statistically related to, the value of an underlying continuous spatial phenomenon $S(x)$ at the corresponding sampling location x . This means that for instance the orders of alpine skis in Lausanne are statistically related to the spatial phenomenon $S(x)$ in Lausanne. Therefore, given the continuous process $S(x)$, the observed data Y are assumed to be independent conditional on $S(x)$ (Diggle et al., 2003, 1998).

In this paper, we consider a Gaussian model with the following specifications:

$$1 \quad E[Y_i | S(\cdot)] = S(x_i) + \sum_{k=1}^p \beta_k d_k(x_i) \quad i = 1, \dots, n \quad p = \# \text{ of covariates}$$

where

$d_k(x_i)$ is the measurement of the k^{th} covariate at the i^{th} location

β_k are the unknown spatial regression parameters

x_i ($Longitude_i, Latitude_i$).

$$2 \quad Var[Y_i | S(\cdot)] = \tau^2 \quad i = 1, \dots, n$$

3 The signal $S(\cdot) = \{S(x): x \in A\}$ is a Gaussian stationary stochastic process with:

$$a \quad E[S(x_i)] = \alpha_0 + \alpha_1 Longitude_i + \alpha_2 Latitude_i + \alpha_3 Longitude_i^2 + \alpha_4 Latitude_i^2 + \alpha_5 Longitude_i x Latitude_i$$

$$b \quad Var[S(x_i)] = \sigma^2$$

$$c \quad Corr[S(x_i), S(x_j)] = \rho(u)$$

where

$u = \|x_i - x_j\|$ is the Euclidian distance between two given locations and

4 The correlation function $\rho(\cdot)$ is specified by the parametric exponential correlation model.

$$\rho(u) = \exp\left(\frac{-u}{\varphi}\right) \quad \varphi > 0.$$

Once the theoretical specifications are made by defining the signal and the correlation function, a crucial step in the data fitting is to find the best model $E[Y_i | S(\cdot)] = S(x_i) + \sum_{k=1}^p \beta_k d_k(x_i)$ for a subset of the available covariates $d_k(x_i)$. To do so, several nested models including different covariates were fitted and compared using the Akaike information criterion (AIC, see Akaike, 1973), according to which the model with the lowest AIC is the best.

For each seasonal product, namely Atomic, Atomic Alpine and Atomic X-Country and each unseasonal product, Wilson, Wilson Racketsports and Wilson Golf, we compare two nested prediction models: M0 (only location data used) and M1 (M0+ weather and socio-economic covariates).

M0 For both seasonal and unseasonal products, in the M0 models, the mean is assumed to be the first order polynomial on the coordinates, since the quadratic terms were not significant.

$$E[Y_i | S(\cdot)] = \alpha_0 + \alpha_1 \text{Longitude}_i + \alpha_2 \text{Latitude}_i$$

M1 In models M1, the expected order value is specified by the coordinates and the covariates, which are:

- number of hotel nights
- percentage of population between 20–64 years old (%)
- private households
- percentage of housing environment and infrastructure surface (%)
- number of companies in the primary sector
- altitude (m)
- snow: thickness of snow measured at 05:40 am (cm)

$$\begin{aligned} E[Y_i | S(\cdot)] = & \alpha_0 + \alpha_1 \text{Longitude}_i + \alpha_2 \text{Latitude}_i + \beta_1 * \text{NbHotelNights} \\ & + \beta_2 * \text{Pop}_{20-64, \text{years}} + \beta_3 * \text{PrivateHouseholds} + \beta_4 * \text{CompPrimSector} \\ & + \beta_5 * \text{HousingSurface} + \beta_6 * \text{Altitude} + \beta_7 * \text{Snow} \end{aligned}$$

Our method has two steps:

- 1 Fitting the geostatistical model for different product groups.
- 2 Using fitted models for demand prediction.

In the first step, we estimate the coefficient parameters that describe the relationship between the response variable and the explanatory variables; and the estimation for parameters that define the covariance structure of the latent process.

The model parameters to be estimated are:

- β the mean parameters
- σ^2 the variance of the signal

τ^2 the variance of the noise

φ the scale parameter of the correlation function.

These parameters are estimated through the maximisation of the log-likelihood using numerical optimisation. The log-likelihood function is obtained from the density of the multivariate Gaussian.

In the second step, the fitted models are used to predict the values of the response variable, namely the order value, at the locations in the studied area where the response variable is unobserved. The prediction of the order value is a linear predictor obtained by minimising the mean squared error prediction under Gaussian modelling assumptions.

Summary of the data analysis process:

- Transformation of the response variable using the Box-Cox transformation in order to meet the assumption of normality.
- Use of the semivariogram for empirically describing the correlation structure according to the separation distance.
- Estimation of the model parameters by maximising the log-likelihood function.
- Use of the fitted model to predict the values of the order value at a given location, by minimising the mean squared error prediction.

The computation is done using the R package *geoR* (Ribeiro and Diggle, 2007).

4 Results

We start with a short spatial exploratory analysis to empirically describe the correlation structure through the semivariogram, followed by the presentation of the model results, the estimations of the parameters. We finally discuss the predictions of the order value for seasonal and unseasonal products all across Switzerland.

4.1 Semivariograms

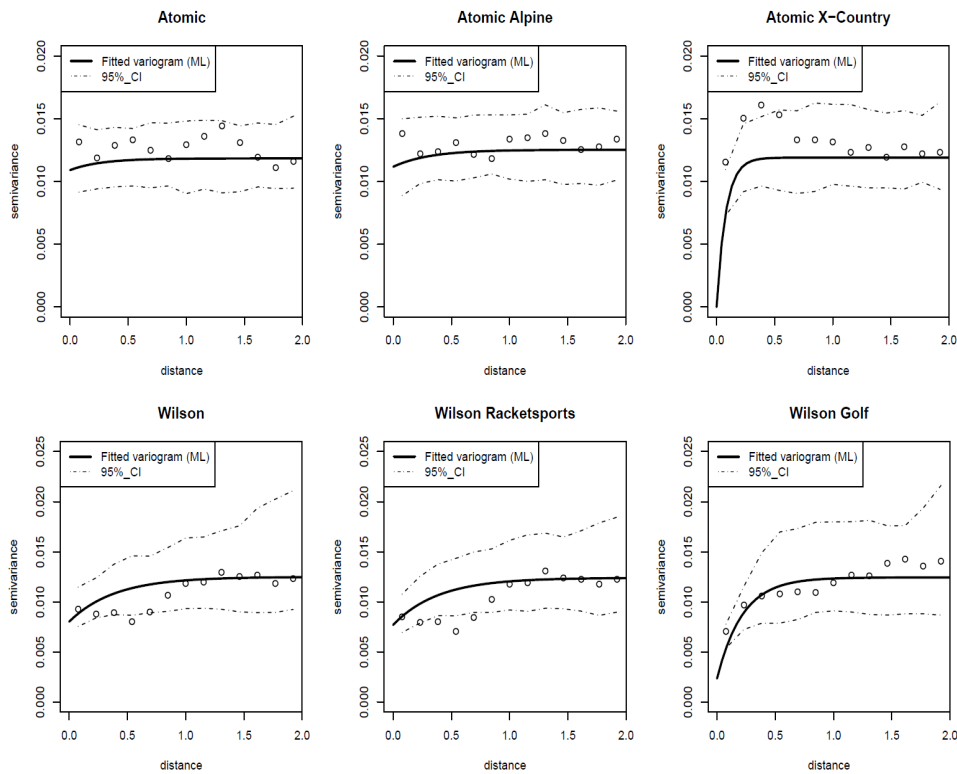
The following semivariograms (Figure 3) provide a descriptive analysis of the spatial association as a function of the separation distance.

The first row of Figure 3 displays the semivariograms for seasonal products and the second row the semivariograms for unseasonal products. The analysis of a semivariogram is made through two main components which are the nugget and the range which is related to the sill, as explained below:

- The nugget is the semivariogram value at the origin ($u = 0$), which should be 0. Having a nugget significantly different from 0 for small distance intervals is named the nugget effect. The nugget effect characterises the eventual discontinuity jump observed at the origin of distances, quantifies the erratic variations of the studied phenomenon, measurements and data errors. The nugget effects of the seasonal products, namely Atomic and Atomic Alpine are significantly different from zero (0.0117 and 0.011), meaning that the seasonality is translated as data errors. The nuggets effects of unseasonal products are lower (0.008, 0.007 and 0.002).

- The sill is the value of the semivariogram at which the semivariogram levels off (theoretical sample variance). The range is the lag distance at which the semivariogram reaches the sill value. The seasonal products have a small range. The fitted semivariogram is flat. That tends to express a spatial independence between the transformed order value and the locations. The unseasonal products have a range parameter around 0.011 and the transformed order value is spatially correlated to the locations.

Figure 3 Empirical semivariogram



Note: Model fitted variogram with 95% confidence bounds.

4.2 Model fitting results

In this section, we present the results of the models defined previously in Section 3. The estimations of the parameters, AICs and p-values are summarised in Tables 2 and 3, for seasonal and unseasonal products, respectively.

According to the AIC, all the models that contain significant socio-economic and/or weather condition covariates (all M1s) are better. These covariates include the number of hotel nights, the percentage of the active population, the number of private households, the number of companies in the primary sector, the percentage of housing environment and infrastructure, altitude and snow. This means that they provide more information about the spatial dependence structure of the transformed order value. The

variance analysis using a chi-squared test shows that the M1 models are significantly different (p-values are all less than $2.2e-16$). These results allow us to conclude that socio-economic and weather conditions have an impact on the spatial demand structure.

Table 2 Model results for seasonal products (see online version for colours)

	<i>Seasonal products</i>					
	<i>Atomic</i>		<i>Atomic Alpine</i>		<i>Atomic X-Country</i>	
	<i>M0</i>	<i>M1</i>	<i>M0</i>	<i>M1</i>	<i>M0</i>	<i>M1</i>
$\hat{\alpha}_0$ [constant]	18.105 (5.52)*	20.747 (4.82)*	9.224 (5.29)	10.804 (5.56)	-21.91 (11.37)	-30.74 (9.98)*
$\hat{\alpha}_1$ [longitude]	0.002 (0.05)	-0.025 (0.05)	-0.01 (0.05)	-0.022 (0.06)	-0.017 (0.11)	-0.026 (0.10)
$\hat{\alpha}_2$ [latitude]	-0.448 (0.12)*	-0.513 (0.11)*	-0.253 (0.11)*	-0.297 (0.12)*	0.41 (0.25)	0.586 (0.22)
$\hat{\beta}_1$ [nb. hotel nights]		-1.117 (0.54)*		-		-
$\hat{\beta}_2$ [pop. 20–64 years]		3.491 (0.4)*		3.124 (0.32)*		-
$\hat{\beta}_3$ [priv. households]		-		-		-
$\hat{\beta}_4$ [comp. prim. sector]		-		-		-
$\hat{\beta}_5$ [% housing surface]		-		-		-
$\hat{\beta}_6$ [altitude]		-		-		3.194 (0.45)*
$\hat{\beta}_7$ [snow]		1.091 (0.33)*		0.921 (0.34)*		2.344 (0.43)*
AIC	-1,522	-1,627	-1,293	-1,214	-1,103	-1,213
p-value		< $2.2e-16$		< $2.2e-16$		< $2.2e-16$

Notes: AIC and estimations of the parameters related to spatial coordinates in light blue, related to the socio-economic covariates in blue and related to weather condition covariates in green. The corresponding standard errors are in parentheses.

*Indicates the significance, with a significance level of 0.05.

At the brand level, the coordinates indicate that the direction of the effect is north-east for all products: the longitude negatively affects the spatial correlation function of all products, while it is the opposite for the latitude.

Seasonal demand and unseasonal demand are not totally impacted by the same socio-economic elements. When comparing Atomic and Wilson, the number of hotel nights and the percentage of the population between 20–64 years old are significant for both seasonal and unseasonal demand. The number of private households per postal code and the number of companies in the primary sector have an impact only on unseasonal demand.

The percentage of housing environment and infrastructure surface per postal code is significant only for Wilson Golf demand. Since golf sport requires a significant infrastructure surface, this result tends to say that the higher the percentage of housing environment and infrastructure surface, the higher the Wilson Golf equipment demand.

Table 3 Model results for unseasonal products (see online version for colours)

	<i>Unseasonal products</i>					
	<i>Wilson</i>		<i>Wilson Racketsports</i>		<i>Wilson Golf</i>	
	<i>M0</i>	<i>M1</i>	<i>M0</i>	<i>M1</i>	<i>M0</i>	<i>M1</i>
$\hat{\alpha}_0$ [constant]	-23.125 (19.97)	-34.243 (6.72)*	-23.34 (18.97)	-34.956 (6.53)*	-21.795 (16.5)	-24.244 (15.59)
$\hat{\alpha}_1$ [longitude]	-0.625 (0.21)*	-0.364 (0.06)*	-0.611 (0.19)*	-0.376 (0.06)*	-0.132 (0.18)	-0.054 (0.18)
$\hat{\alpha}_2$ [latitude]	0.512 (0.43)	0.68 (0.15)*	0.515 (0.41)	0.7 (0.14)*	0.429 (0.35)	0.459 (0.33)*
$\hat{\beta}_1$ [nb. hotel nights]		-7.619 (1.88)*		-7.435 (1.84)*		-
$\hat{\beta}_2$ [pop. 20–64 years]		7.139 (0.79)*		6.262 (0.69)*		-
$\hat{\beta}_3$ [priv. households]		5.747 (2.04)*		5.661 (2.03)*		-
$\hat{\beta}_4$ [comp. prim. sector]		-0.981 (0.49)*		-1.01 (0.5)*		-
$\hat{\beta}_5$ [% housing surface]		-		-		1.961 (0.56)*
$\hat{\beta}_6$ [altitude]		2.521 (0.81)*		2.577 (0.79)*		2.683 (0.65)*
$\hat{\beta}_7$ [snow]		-		-		-
AIC	-1,911	-2,037	-1,861	-1,981	-1,123	-1,240
p-value		< 2.2e-16		< 2.2e-16		< 2.2e-16

Notes: AIC and estimations of the parameters related to spatial coordinates in light blue, related to the socio-economic covariates in blue and related to weather condition covariates in green. The corresponding standard errors are in parentheses.

*Indicates the significance, with a significance level of 0.05.

The annual number of registered hotel nights in the location impacts the demand for seasonal products more negatively than the demand for unseasonal products, at the brand level. At the product or product family level, the annual number of registered hotel nights negatively affects only the Wilson Racketsports demand which corresponds to more familiar sports (tennis, badminton and squash). This result could reflect the fact that sport equipment is usually bought where one lives and rarely while travelling. Therefore, the more the registered hotel nights, relatively the lower the sport equipment demand.

At the brand level, the percentage of labour force positively affects both seasonal and unseasonal demand. Indeed, since the labour force represents the purchasing power, an increase in the population between 20–64 years old tends to increase the sport equipment demand for both seasonal and unseasonal products. But at the product family level, the positive effect of the labour force is significant for Atomic Alpine and Wilson Racketsports and is not significant for Alpine X-Country and Wilson Golf demand.

The number of companies in the primary sector and the percentage of housing and infrastructure area reflect the distinction between big cities and the countryside. They are significant only for some unseasonal products, namely Wilson Racketsports and not for Wilson Golf demand. This result tends to indicate that Wilson Racketsports equipment, used for familiar sports, is bought more in big cities.

The yearly maximum registered snow is significant only for seasonal products and positively affects seasonal demand. The altitude does not affect Atomic and Atomic Alpine demand and positively affects unseasonal demand and Atomic X-Country demand. This could indicate that for seasonal demand, people tend to bring their own winter sport equipment from home and tend not to buy it in ski resorts or other winter resorts. On the contrary, unseasonal sport equipment is bought everywhere.

These results allow us to conclude that socio-economic and weather conditions have an impact on the spatial demand structure and provide more information about this structure for both seasonal and unseasonal products.

4.3 Prediction results

We developed the spatial interpolation, called kriging, for the geographical area of Switzerland, ordinary kriging is an estimation procedure that uses weighted averages specified through the variogram model that describes the spatial correlation. In other words, it consists of using a weighted average of neighbouring samples to estimate the unknown value at a given location. These weights are optimised using the variogram model and the location of the samples.

In Figure 4, we show the predicted order values for both seasonal (first column) and unseasonal products (second column) obtained by using the corresponding model M1 that contains socio-economic features and weather conditions.

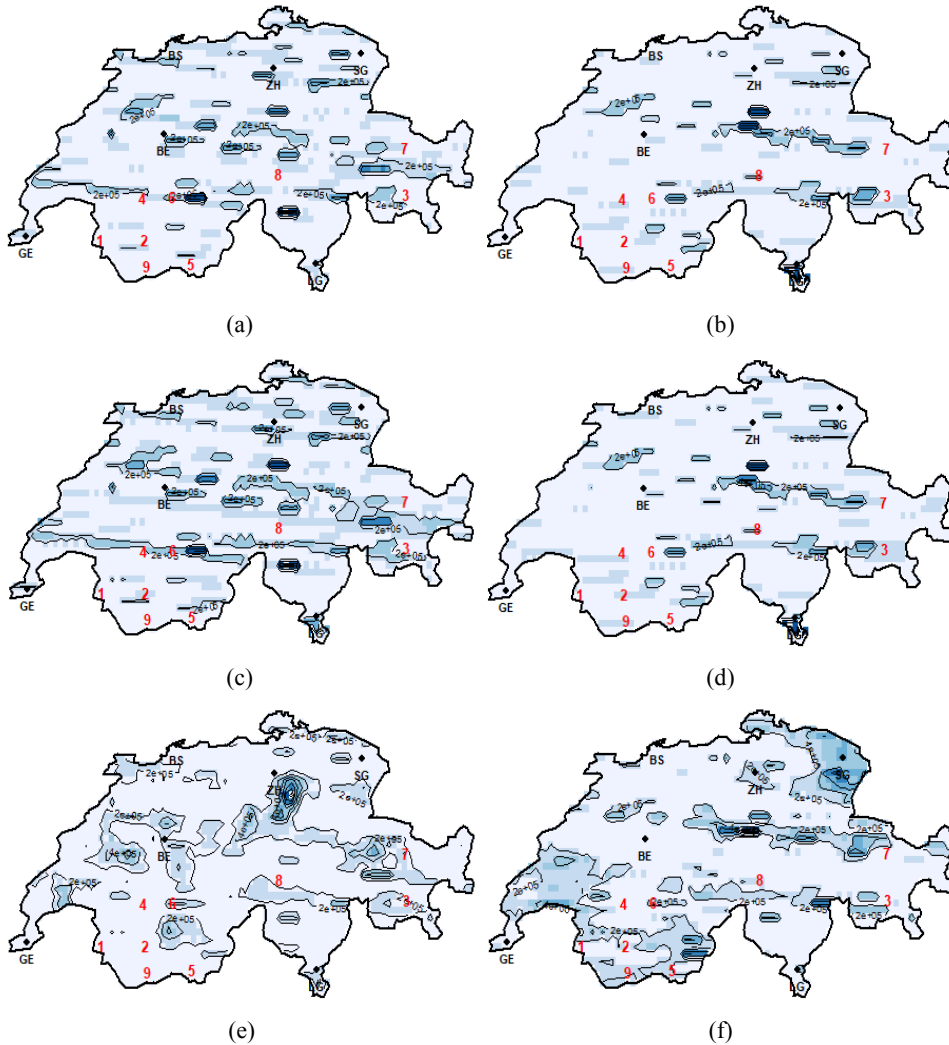
The first line [Figures 4(a) and 4(b)] corresponds to the predictions of the order value for seasonal and unseasonal products at the brand level, namely Atomic and Wilson, respectively. The two graphs are quite different. For Atomic [Figure 4(a)], the prediction of the order value tends to be high around big cities such as Bern, Zurich and St. Gallen and around big ski resorts such as Davos, St. Moritz and Andermatt. The map of Wilson predictions [Figure 4(b)] is relatively uniform everywhere across Switzerland.

Figures 4(c) and 4(e) display the predictions of the order value for seasonal products at the product level, namely Atomic Alpine and Atomic X-Country, respectively. The map corresponding to Atomic Alpine demand [Figure 4(c)] suggests that the order value is high around big cities such as Bern, Zurich and St. Gallen. The demand is also high in the region of Grison, where there are famous ski resorts such as Davos, St. Moritz or in the north of Ticino (Andermatt area), where the demand is predicted to be high. Figure 4(e) corresponding to the map of Atomic X-Country demand suggests that the order value is high around Zurich, Bern and in the region of Grison. This finding could help the company to decide about the location of a new shop selling the considered seasonal product.

The order values for unseasonal products at the product level are displayed in Figures 4(d) and 4(f), corresponding to Wilson Racketsports and Wilson Golf, respectively. The demand levels for Wilson Racketsports, which are considered as the most popular sports, are predicted to be relatively regular everywhere across the country [Figure 4(d)]. There are some regions where these levels are slightly higher, such as around Zurich and in the Grison around St. Moritz. There is also a high demand in

Lugano, in the south of Ticino. The predictions are more irregular for Wilson Golf equipment demand [Figure 4(f)]. The order value is higher in the region of St. Gallen, Luzern and in the Valais.

Figure 4 Map showing the prediction of the order value in CHF for the seasonal demand of Atomic across Switzerland (kriging according to M1), (a) Atomic (b) Wilson (c) Atomic Alpine (d) Wilson Racketsports (e) Atomic X-Country (f) Wilson Golf (see online version for colours)

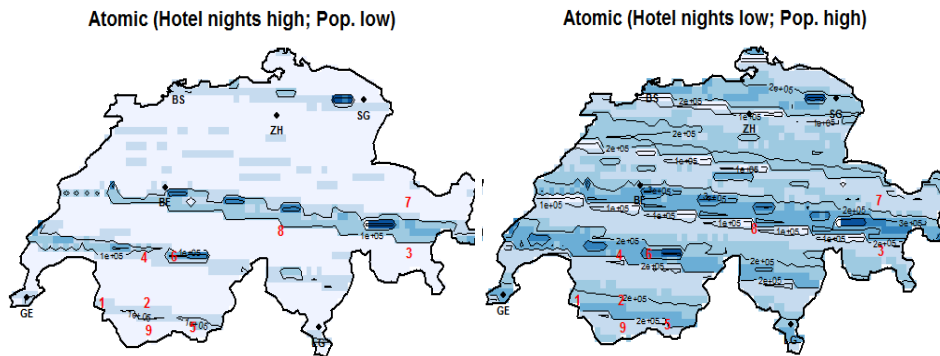


Note: With cities [Geneva (GE), Bern (BE), Lugano (LG), Zurich (ZH), Basel (BS) and St. Gallen (SG)] and ski resorts [Champéry (1), Nendaz (2), St. Moritz (3), Gstaad (4), Zermatt (5), Adelboden (6), Davos (7), Andermatt (8) and Verbier (9)].

In order to analyse the effect of a variable, we first set this variable at a certain level (low or high), and then we predict the order value and compare them obtained maps. The low and high levels correspond to the 0.25 and 0.8 sample quantiles of observed data.

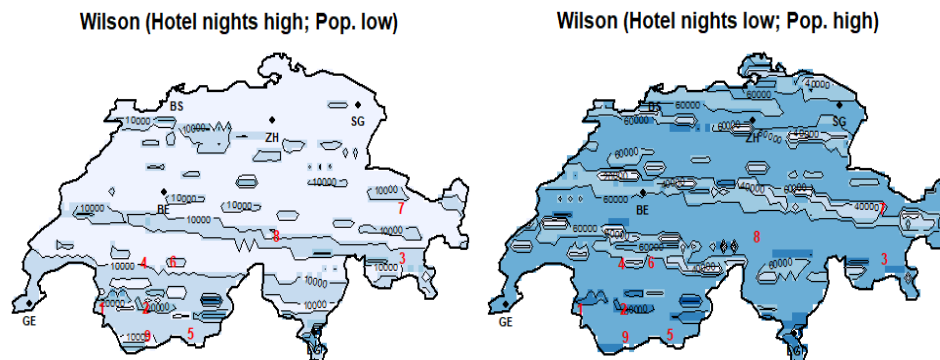
Figures 5 and 6 provide prediction maps of the order value in Swiss francs (CHF) for seasonal and unseasonal brands, respectively Atomic and Wilson, using the corresponding model M1 different levels of two covariates, namely the number of hotel nights and the proportion of the labour population. Since the effect of the number of hotel nights on the order value is negative (Table 4), we then start with the high level of number of hotel nights. We have the opposite for the proportion of the labour population, meaning that since its effect is positive, we set the first level to be low. The effects of these covariates on the Atomic products (Figure 5) show that the order value is higher and especially around ski resorts. The predicted order value of Wilson is higher but more uniform across the country (Figure 6).

Figure 5 Map showing the prediction of the order value in CHF for the seasonal demand of Atomic across Switzerland (kriging according to M1) (see online version for colours)



Note: With cities [Geneva (GE), Bern (BE), Lugano (LG), Zurich (ZH), Basel (BS) and St. Gallen (SG)] and ski resorts [Champéry (1), Nendaz (2), St. Moritz (3), Gstaad (4), Zermatt (5), Adelboden (6), Davos (7), Andermatt (8) and Verbier (9)].

Figure 6 Map showing the prediction of the order value in CHF for the unseasonal demand of Wilson across Switzerland (kriging according to M1) (see online version for colours)

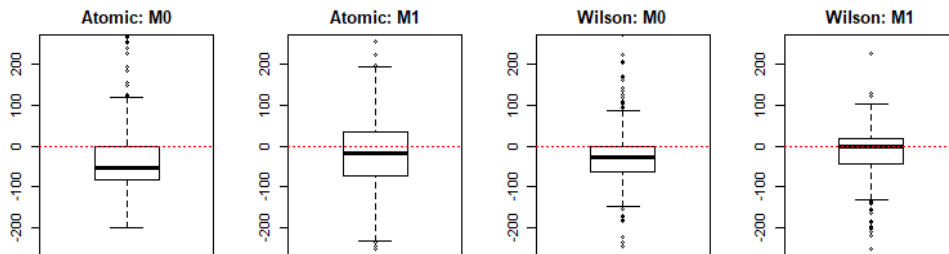


Note: With cities [Geneva (GE), Bern (BE), Lugano (LG), Zurich (ZH), Basel (BS) and St. Gallen (SG)] and ski resorts [Champéry (1), Nendaz (2), St. Moritz (3), Gstaad (4), Zermatt (5), Adelboden (6), Davos (7), Andermatt (8) and Verbier (9)].

Table 4 Summary of covariate effects according to brand seasonality (see online version for colours)

Covariate\brand seasonality	Seasonal	Unseasonal
Nb. hotel nights	↘	↘
Pop. 20–64 years	↗	↗
Priv. households	-	↗
Comp. prim. sector	-	↘
% housing surface	-	-
Altitude	-	↗
Snow	↗	-

The model validation is performed by comparing observed and predicted values. This is done by *leaving-one-out* cross-validation, consisting of removing one point from the dataset and predicting it using the remaining points. This process is done for each point. Figure 7 displays the boxplots of the errors provided by the cross-validation analyses. These graphs show that the models containing only the coordinates (M0s) tend to underestimate the demand level. This is the case for seasonal and unseasonal products. The M1 boxplots, that is, the models that contain not only the coordinates, but also socio-economic features and weather conditions, show a clear improvement in the demand forecasting accuracy. The reduction of the median absolute percentage error (*MdAPE*) is -25% , -31% and -26% for seasonal products (respectively Atomic, Atomic Alpine and Atomic X-Country) and -18% , -11% and -48% for unseasonal products (respectively Wilson, Wilson Racketsports and Wilson Golf).

Figure 7 Boxplots of the prediction errors for different models (see online version for colours)

5 Conclusions

In this paper, we studied how seasonal and unseasonal demands spatially vary according to socio-economic aspects and weather conditions, and how this additional information could be used to improve the accuracy of seasonal demand forecasting. We analysed real business data for orders in Switzerland along with socio-economic features and weather conditions. We found that the effect of the year on the spatial trend is not significant. After removing the trend, the studied spatial structure does not significantly differ from

one year to another. This means, we can generalise the results among different years. By analysing the most recent year using model-based geostatistics, we found that the incorporation of socio-economic data and weather conditions in the model provides more information about the spatial dependence structure of the demand for seasonal and unseasonal products than the tested pure geostatistical model does. Moreover, the incorporation of this additional information increases the demand forecasting accuracy. We discovered that the analyses according to the spatial coordinates tend to systematically underestimate the order value. This bias is corrected in the second model containing more information about the socio-economic environment and weather conditions. The incremental improvement corresponds to a reduction in mean squared errors by -25% , -31% , -26% -18% , -11% and -48% , for, respectively, Atomic, Atomic Alpine, Atomic X-Country, Wilson, Wilson Racketsports products and Wilson Golf. In addition, the kriging method, commonly used in geostatistics, provides the prediction maps of the order value across Switzerland, which is helpful, for example, when deciding on the location of a new shop.

References

- Akaike, H. (1973) 'Maximum likelihood identification of Gaussian autoregressive moving average models', *Biometrika*, Vol. 60, No. 2, pp.255–265.
- Benndorf, J. (2014) 'Moving towards real-time management of mineral reserves – a geostatistical and mine optimization closed-loop framework', in *Mine Planning and Equipment Selection*, pp.989–999, Springer International Publishing, Cham.
- Box, G.E. and Cox, D.R. (1964) 'An analysis of transformations', *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 26, No. 2, pp.211–252.
- Caeiro, S., Goovaerts, P., Painho, M. and Costa, M.H. (2003) 'Delineation of estuarine management areas using multivariate geostatistics: the case of Sado Estuary', *Environmental Science & Technology*, Vol. 37, No. 18, pp.4052–4059.
- Diggle, P.J., Ribeiro, P.J. and Christensen, O.F. (2003) 'An introduction to model-based geostatistics', in *Spatial Statistics and Computational Methods*, pp.43–86, Springer, New York, NY.
- Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998) 'Model-based geostatistics', *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Vol. 47, No. 3, pp.299–350.
- Fadiga, M.L., Misra, S.K. and Ramirez, O.A. (2005) 'US consumer purchasing decisions and demand for apparel', *Journal of Fashion Marketing and Management: An International Journal*, Vol. 9, No. 4, pp.367–379.
- Gage, E. and Cooper, D.J. (2015) 'The influence of land cover, vertical structure, and socioeconomic factors on outdoor water use in a Western US city', *Water Resources Management*, Vol. 29, No. 10, pp.3877–3890.
- Gomes, V.A., Pitombo, C.S., Rocha, S.S. and Salgueiro, A.R. (2016) 'Kriging geostatistical methods for travel mode choice: a spatial data analysis to travel demand forecasting', *Open Journal of Statistics*, Vol. 6, No. 3, p.514.
- Inamura, T., Goto, K., Iida, M., Nonami, K., Inoue, H. and Umeda, M. (2004) 'Geostatistical analysis of yield, soil properties and crop management practices in paddy rice fields', *Plant Production Science*, Vol. 7, No. 2, pp.230–239.
- Jain, A. and Ormsbee, L.E. (2002) 'Short-term water demand forecast modeling techniques – conventional methods versus AI', *Journal (American Water Works Association)*, Vol. 94, No. 7, pp.64–72.

- Krige, D.G. (1951) 'A statistical approach to some basic mine valuation problems on the Witwatersrand', *Journal of the Southern African Institute of Mining and Metallurgy*, Vol. 52, No. 6, pp.119–139.
- Legleiter, C.J. and Kyriakidis, P.C. (2008) 'Spatial prediction of river channel topography by kriging', *Earth Surface Processes and Landforms*, Vol. 33, No. 6, pp.841–867.
- Lustgarten, A (2005) 'Getting ahead of the weather', *Fortune*, 21 February, EBSCOhost, Europe.
- Matheron, G. (1963) 'Principles of geostatistics', *Economic Geology*, Vol. 58, No. 8, pp.1246–1266.
- McBratney, A.B. (1992) 'On variation, uncertainty and informatics in environmental soil management', *Soil Research*, Vol. 30, No. 6, pp.913–935.
- Moral, F.J., Terrón, J.M. and Da Silva, J.M. (2010) 'Delineation of management zones using mobile measurements of soil apparent electrical conductivity and multivariate geostatistical techniques', *Soil and Tillage Research*, Vol. 106, No. 2, pp.335–343.
- Moral, F.J., Terrón, J.M. and Rebollo, F.J. (2011) 'Site-specific management zones based on the Rasch model and geostatistical techniques', *Computers and Electronics in Agriculture*, Vol. 75, No. 2, pp.223–230.
- Morari, F., Castrignanò, A. and Pagliarin, C. (2009) 'Application of multivariate geostatistics in delineating management zones within a gravelly vineyard using geo-electrical sensors', *Computers and Electronics in Agriculture*, Vol. 68, No. 1, pp.97–107.
- Muthuwatta, L.P., Bos, M.G. and Rientjes, T.H.M. (2010) 'Assessment of water availability and consumption in the Karkheh River Basin, Iran – using remote sensing and geo-statistics', *Water Resources Management*, Vol. 24, No. 3, pp.459–484.
- Namysłowska-Wilczyńska, B. and Wilczyński, A. (2010) '3D electric power demand forecasting as a tool for planning electrical power firm's activity by means of geostatistical methods', *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu. Ekonometria*, Vol. 28, No. 91 Prognostowanie = Forecasting, pp.95–112.
- Namysłowska-Wilczyńska, B. and Wilczyński, A. (2015) 'Geostatistical characteristics of the structure of spatial variation of electrical power in the national 110 KV network including results of variogram model components filtering', *Acta Energetica*, Vol. 22, No. 1, pp.72–87.
- Nelson, M.R., Orum, T.V., Jaime-Garcia, R. and Nadeem, A. (1999) 'Applications of geographic information systems and geostatistics in plant disease epidemiology and management', *Plant Disease*, Vol. 83, No. 4, pp.308–319.
- Oliver, M.A. (2010) 'An overview of geostatistics and precision agriculture', in *Geostatistical Applications for Precision Agriculture*, pp.1–34, Springer, Netherlands.
- Peng, Y.H., Hsu, C.S. and Huang, P.C. (2015) 'An investigation of spacial approaches for crop price forecasting in different Taiwan markets', in *2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, IEEE, November, pp.176–179.
- Petersen, D.L., Erickson, M.L., Roberson, R. and Siekmeier, J. (2007) 'Intelligent soil compaction: Geostatistical data analysis and construction specifications', in *Transportation Research Board 86th Annual Meeting*, No. 07-2858.
- Polebitski, A.S. and Palmer, R.N. (2009) 'Seasonal residential water demand forecasting for census tracts', *Journal of Water Resources Planning and Management*, Vol. 136, No. 1, pp.27–36.
- Prasetyowati, S.S., Imrona, M., Ummah, I. and Sibaroni, Y. (2016) 'Prediction of public transportation occupation based on several crowd spots using ordinary kriging method', *Journal of Innovative Technology and Education*, Vol. 3, No. 1, pp.93–104.
- Pringle, M.J., Cook, S.E. and McBratney, A.B. (2004a) 'Field-scale experiments for site-specific crop management. Part I: design considerations', *Precision Agriculture*, Vol. 5, No. 6, pp.617–624.
- Pringle, M.J., McBratney, A.B. and Cook, S.E. (2004b) 'Field-scale experiments for site-specific crop management. Part II: a geostatistical analysis', *Precision Agriculture*, Vol. 5, No. 6, pp.625–645.

- Regnier, E. (2008) 'Doing something about the weather', *Omega*, Vol. 36, No. 1, pp.22–32.
- Rendu, J.M. (2002) 'Geostatistical simulations for risk assessment and decision making: the mining industry perspective', *International Journal of Surface Mining, Reclamation and Environment*, Vol. 16, No. 2, pp.122–133.
- Ribeiro, P. and Diggle, P. (2007) 'The geoR package', *R News*, Vol. 1, No. 2, pp.14–18.
- Sciarretta, A., Girma, M., Tikubet, G., Belayehun, L., Ballo, S. and Baumgärtner, J. (2005) 'Development of an adaptive tsetse population management scheme for the Luke community, Ethiopia', *Journal of Medical Entomology*, Vol. 42, No. 6, pp.1006–1019.
- Srivastava, R.M. (2013) 'Geostatistics: a toolkit for data analysis, spatial prediction and risk management in the coal industry', *International Journal of Coal Geology*, Vol. 112, pp.2–13.
- Stelzenmüller, V., Maynou, F., Bernard, G., Cadiou, G., Camilleri, M., Crec'hriou, R., Lenfant, P. et al. (2008) 'Spatial assessment of fishing effort around European marine reserves: implications for successful fisheries management', *Marine Pollution Bulletin*, Vol. 56, No. 12, pp.2018–2026.
- Tuominen, S., Fish, S. and Poso, S. (2003) 'Combining remote sensing, data from earlier inventories, and geostatistical interpolation in multisource forest inventory', *Canadian Journal of Forest Research*, Vol. 33, No. 4, pp.624–634.
- Warner, J.W., Tamayo-Lara, C., Khazaei, E. and Manghi, F. (2006) 'Stochastic management modeling of a pump and treat system at the Rocky Mountain Arsenal near Denver, Colorado', *Journal of Hydrology*, Vol. 328, No. 3, pp.523–537.
- Zhang, Y., Xian, C., Chen, H., Grieneisen, M.L., Liu, J. and Zhang, M. (2016) 'Spatial interpolation of river channel topography using the shortest temporal distance', *Journal of Hydrology*, Vol. 542, pp.450–462.

Notes

$$1 \quad MdAPE = med \left[\frac{(Forecast_t - Actual_t) * 100}{Actual_t} \right].$$

Part II: Non-linear causal inference

Chapter 5

Causal discovery for heteroscedastic financial series

Abstract

Inferring causality between financial assets is a common and fundamental subject in finance. The widely used Granger causality allows to determine whether one time series is useful in forecasting another. Under Granger causality, the cause happens prior to its effect. In this paper, we propose a new method to understand intrinsic causal mechanism between series, unconditionally on time. Dealing with heteroskedastic financial data, we investigate causal relations not only in mean but from the perspective of location, scale and shape parameters of the underlying distribution. The proposed two-steps method relies on bayesian additive models for location, scale and shape (BAMLSS) and on causal additive models (CAM), admitting non-linear and non-gaussian causal multiplicative noise models. Based on an extensive simulation study, our approach globally outperforms standard causal discovery methods of data science. When applied to financial indices, we find evidence of an un-lagged causal effect of the shares on the index they compose. We detail the methodology for the bivariate case but, in the empirical study, we show how to extend the causal discovery to the multivariate case.

5.1 Introduction

A key challenge in finance is to understand relationship between financial assets for both investment and risk management purposes. Granger causality [Granger and Morgenstern, 1963] analysis is the standard method for achieving this. It provides information about the dynamic interactions between time series, conditionally on time. More precisely, a time series X is said to Granger-cause Y if it can be tested that lagged values of X provide statistically significant information about future values of Y . There are two serious limitations of this approach. First, the causality is inferred at the level of the mean only and not at other distribution characteristics. To remedy this, Chuang et al. [2009] investigated Granger causal relations from the perspective of conditional quantiles. In this paper, motivated by the heteroscedastic stylised feature of financial time series, we infer causal relations at higher moments of the data distribution. Second, the autoregressive structure underlying Granger causal-

ity is restrictive. We relax any parametric assumption about the functional form of the causal effect. Our approach offers a new methodology to detect intrinsic causal mechanism between financial series, unconditionally on time. It is based on recent data science developments inferring causal relationships from observational data using conditional independence [Rubin, 1974; Pearl, 2009; Spirtes and Zhang, 2016].

Consider two random variables X and Y which can represent financial time series and that are linked by an intrinsic causal mechanism. We assume that there is no latent confounding variable causing both X and Y . If X is the cause of Y we note $X \rightarrow Y$ and the traditional structural equation model approach is written as

$$Y = f(X, \varepsilon), \quad (5.1)$$

where f is a causal mechanism linking Y to its direct cause X , and ε the noise variable [Galles and Pearl, 1997].

Many papers have studied additive noise models (ANM) [Peters et al., 2014; Hoyer et al., 2009] for which

$$Y = f(X) + \varepsilon. \quad (5.2)$$

To allow heteroscedasticity, we consider in this paper a multiplicative model with respect to the noise variable, that is

$$Y = f(X) + g(X)\varepsilon, \quad (5.3)$$

where the noise is $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 > 0$ and $f(\cdot)$ and $g(\cdot)$ are arbitrary functions.

Causality identification in the bivariate case is challenging [Galles and Pearl, 1995]. In the additive context, certain approaches impose model specifications or restrictions [Nowzohour and Bühlmann, 2016; Hoyer et al., 2009]. For linear causal models ($Y = bX + \varepsilon$), if X and ε are gaussian, the causal direction is identifiable, due to the independent component analysis (ICA) theory [Hyvärinen et al., 2004]. The linear non-Gaussian causal model, known as LinGAM [Shimizu et al., 2006] also relies on ICA with the additional assumption that disturbance variables have non-Gaussian distributions of non-zero variances. Although linear causal models with additive noise are often used because they are well understood, in reality, many causal relationships are far from being linear. Non-linearities in the data-generating process provide more information on the underlying causal system since these models allow more aspects of the true data-generating mechanisms to be identified ($Y = f(X) + \varepsilon$) [Hoyer et al., 2009]. The post-nonlinear causal model ($Y = g(f(X) + \varepsilon)$) provided by Zhang and Hyvärinen [2009] aims to distinguish the cause from effect by analyzing the non-linear effect of the cause, the inner noise effect, and the measurement distortion effect in the observed variables.

In this paper, we address the case of non-linear and non-gaussian causal multiplicative noise model ($Y = f(X) + g(X)\varepsilon$) [Tagasovska et al., 2018; Goudet et al., 2018]. To distinguish the cause from the effect, we use a two-step method based on bayesian additive models for location, scale and shape parameters and on causal additive models applied to the fitted estimated parameters. In other words we first consider the two multiplicative noise models

$Y = f_y(X) + g_y(X)\varepsilon_y$ (respectively $X = f_x(Y) + g_x(Y)\varepsilon_x$), and using BAMLSS, we compute the resulting fitted location, scale and shape of the estimated parameters $\hat{\beta}_{Y|X}$ (respectively $\hat{\beta}_{X|Y}$). Second, to discover causality, we apply the bivariate causal additive model (BiCAM) of Peters et al. [2014]. The BiCAM approach has been proven to be an efficient and computationally fast method to discover causality in the Gaussian context. Relying on the asymptotic normality of the BAMLSS estimator of β we apply BiCAM on pairwise fitted parameters of $\hat{\beta}_{Y|X}$ and $\hat{\beta}_{X|Y}$. We provide a rule based on the BiCAM's scores of each parameter to identify the causal direction.

The rest of the paper is organized as follows: in Section 5.2, we detail our two step method, and in Section 5.3 we run an extensive simulation study to assess the accuracy of our method. Intrinsic causal discovery between pairs of indices and shares are presented in the first part of Section 5.4, the second part describing the extension to the multivariate setting of several shares. We conclude in Section 5.5.

5.2 Causal discovery for heteroscedastic model

In this section we present our two-step method for causal heteroscedastic model (CHM), which consists of applying a BiCAM method on the fitted values of the estimated BAMLSS models parameters.

5.2.1 First step: BAMLSS

Bayesian additive models for location, scale and shape [Umlauf et al., 2017] are bayesian version of generalized additive models for location, scale and shape (GAMLSS) [Rigby and Stasinopoulos, 2005; Stasinopoulos and Rigby, 2007] that relaxe the distributional assumptions of the response variable and allow the modelization of the mean (location) and higher moments (usually scale and shape) using covariates. Each parameter of the distribution is linked to an additive predictor as for generalized additive models (GAM) [Hastie and Tibshirani, 1990] and the covariate effects can have flexible forms such as, for example linear, non-linear, spatial or random effects. BAMLSS handles complex models, like for instance considering a response distribution out of the exponential family or when multiple predictors contain several smooth effects. In these cases, the bayesian approach that uses Markov chain Monte Carlo (MCMC) simulation techniques provides valid and credible confidence intervals while standard confidence intervals based on asymptotic properties of the maximum likelihood estimators fail.

Model structure

BAMLSS models are based on n observations and assume conditional independence of individual response observations y_i given a set of covariates \mathbf{W} , (i.e: $y_i | \mathbf{W} \perp y_j | \mathbf{W}, \forall i \neq j; i, j = 1, \dots, n$). All parameters of the response distribution can be modeled using explanatory variables such that

$$Y \sim \mathcal{D}_Y(h_1(\theta_1) = \eta_1, h_2(\theta_2) = \eta_2, \dots, h_K(\theta_K) = \eta_K)$$

where \mathcal{D}_Y denotes a parametric distribution for the response variable Y with K parameters $\theta_k, \theta_k \in \Gamma_\theta, k = 1, \dots, K$, that are linked to additive predictors using known monotonic and twice differentiable functions $h_k(\cdot)$. For example for the gaussian distribution we have $Y \sim \mathcal{N}(h_\mu(\mu) = \eta_\mu, h_{\sigma^2}(\sigma^2) = \eta_{\sigma^2})$. The k -th additive predictor is given by

$$h_k(\theta_k) = \eta_k := \eta_k(\mathbf{W}_k; \boldsymbol{\beta}_k) = \sum_{j=1}^{J_k} f_{jk}(\mathbf{W}_{jk}; \boldsymbol{\beta}_{jk})$$

where:

- $\mathbf{W}_k = (\mathbf{W}_{1k}; \dots; \mathbf{W}_{J_k k})^T$ is the combined design matrix for the k -th parameter
- $f_{jk}(\cdot)$ are unspecified and possibly non-linear functions of subvectors of \mathbf{W} collecting all available covariate information, $j = 1, \dots, J_k$ and $k = 1, \dots, K$.
- $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_{1k}, \dots, \boldsymbol{\beta}_{J_k k})^T, \boldsymbol{\beta}_{J_k k} \in \Gamma_\beta$, are coefficients that need to be estimated from the data.

Model fitting and inference

The estimation of the probability density function $d_Y(\mathbf{y}|\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)$ requires to evaluate the log-likelihood function:

$$l(\boldsymbol{\beta}; \mathbf{y}, \mathbf{W}) = \sum_{i=1}^n \log d_y(y_i; \theta_{i1} = h_1^{-1}(\eta_{i1}(\mathbf{w}_i; \boldsymbol{\beta}_1)), \dots, \theta_{iK} = h_K^{-1}(\eta_{iK}(\mathbf{w}_i; \boldsymbol{\beta}_K))).$$

By assigning prior distributions $p_{jk}(\cdot)$ to the individual model component we obtain the log-posterior:

$$\log \pi(\boldsymbol{\beta}, \boldsymbol{\tau}; \mathbf{y}, \mathbf{W}, \boldsymbol{\alpha}) \propto l(\boldsymbol{\beta}; \mathbf{y}, \mathbf{W}) + \sum_{k=1}^K \sum_{j=1}^{J_k} \log p_{jk}(\boldsymbol{\beta}_{jk}; \boldsymbol{\tau}_{jk}, \boldsymbol{\alpha}_{jk})$$

where $\boldsymbol{\tau} = (\boldsymbol{\tau}_1^T, \dots, \boldsymbol{\tau}_K^T)^T$ with $\boldsymbol{\tau}_{jk} \in \Gamma_\tau$ is the vector of all assigned hyper-parameters used within prior functions $p_{jk}(\cdot)$ and similarly $\boldsymbol{\alpha}$ is the set of all fixed prior specifications. The rather general prior for the jk -th model term is given by

$$p_{jk}(\boldsymbol{\beta}_{jk}; \boldsymbol{\tau}_{jk}, \boldsymbol{\alpha}_{jk}) \propto d_{\boldsymbol{\beta}_{jk}}(\boldsymbol{\beta}_{jk}|\boldsymbol{\tau}_{jk}; \boldsymbol{\alpha}_{\boldsymbol{\beta}_{jk}}) \cdot d_{\boldsymbol{\tau}_{jk}}(\boldsymbol{\tau}_{jk}|\boldsymbol{\alpha}_{\boldsymbol{\beta}_{jk}})$$

with prior densities or combinations of densities $d_{\boldsymbol{\beta}_{jk}}(\cdot)$ and $d_{\boldsymbol{\tau}_{jk}}(\cdot)$ that depend on the type of the covariate and prior assumptions about $f_{jk}(\cdot)$.

Bayesian point estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ for the posterior mean estimation is obtained by solving high dimensional integrals given by

$$E(\boldsymbol{\beta}, \boldsymbol{\tau}|\mathbf{y}, \mathbf{W}, \boldsymbol{\alpha}) = \int_{\Gamma_\beta} \int_{\Gamma_\tau} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\tau} \end{pmatrix} \pi(\boldsymbol{\beta}, \boldsymbol{\tau}; \mathbf{y}, \mathbf{W}, \boldsymbol{\alpha}) d \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\tau} \end{pmatrix}$$

which can be rarely solved analytically and therefore need to be approximated by numerical techniques such as MCMC simulations through iterative algorithm with an updating scheme of type

$$(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\tau}^{(t+1)}) = U(\boldsymbol{\beta}^{(t)}, \boldsymbol{\tau}^{(t)}; \mathbf{y}, \mathbf{W}, \boldsymbol{\alpha})$$

at step $t + 1$ where $U(\cdot)$ is an updating function. MCMC samples for the regression coefficients $\boldsymbol{\beta}_{jk}$ can be derived by each of the following methods namely Random-walk Metropolis [Metropolis et al., 1953; Gelman et al., 1996], Derivative-based Metropolis-Hastings [Hastings, 1970] and Slice sampling [Neal, 2003]. Consider for instance the Derivative-based Metropolis-Hastings method.

The sampler proceeds by drawing a candidate $\boldsymbol{\beta}_{jk}^*$ from a symmetric jumping distribution $q(\boldsymbol{\beta}_{jk}^* | \boldsymbol{\beta}_{jk}^{(t)})$ which is commonly a normal distribution $\mathcal{N}(\mu_{jk}^{(t)}, \sum_{jk}^{(t)})$ centered at the current iterate with:

- $\mu_{jk}^{(t)} = \boldsymbol{\beta}_{jk}^{(t)} - [\mathbf{J}_{kk}(\boldsymbol{\beta}_{jk}^{(t)}) + \mathbf{G}_{jk}(\boldsymbol{\tau}_{jk})]^{-1} \mathbf{s}(\boldsymbol{\beta}_{jk}^{(t)})$
- $(\sum_{jk}^{(t)})^{-1} = -\mathbf{H}_{kk}(\boldsymbol{\beta}_{jk}^{(t)})$
- $\boldsymbol{\beta}_{jk}^{(t+1)} = \boldsymbol{\beta}_{jk}^{(t)} - [\mathbf{J}_{kk}(\boldsymbol{\beta}_{jk}^{(t)}) + \mathbf{G}_{jk}(\boldsymbol{\tau}_{jk})]^{-1} \mathbf{s}(\boldsymbol{\beta}_{jk}^{(t)})$

where the score vector $\mathbf{s}(\cdot)$ and the Hessian matrix are explicated bellow:

$$\begin{aligned} \mathbf{s}(\boldsymbol{\beta}) &= \frac{\partial \log \pi(\boldsymbol{\beta}, \boldsymbol{\tau}; \mathbf{y}, \mathbf{W}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} = \frac{\partial l(\boldsymbol{\beta}; \mathbf{y}, \mathbf{W})}{\partial \boldsymbol{\beta}} + \sum_{k=1}^K \sum_{j=1}^{J_k} \frac{\partial \log p_{jk}(\boldsymbol{\beta}_{jk}; \boldsymbol{\tau}_{jk}, \boldsymbol{\alpha}_{jk})}{\partial \boldsymbol{\beta}} \\ \mathbf{H}_{jk}(\boldsymbol{\beta}) &= \frac{\partial \mathbf{s}(\boldsymbol{\beta}_j)}{\partial \boldsymbol{\beta}_k} = \frac{\partial^2 \log \pi(\boldsymbol{\beta}, \boldsymbol{\tau}; \mathbf{y}, \mathbf{W}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_k^T} \\ \mathbf{J}_{jk}(\boldsymbol{\beta}) &= \frac{\partial^2 l(\boldsymbol{\beta}; \mathbf{y}, \mathbf{W})}{\partial \boldsymbol{\beta}_{jk} \partial \boldsymbol{\beta}_{jk}^T} \\ \mathbf{G}_{jk}(\boldsymbol{\tau}_{jk}) &= \frac{\partial^2 \log p_{jk}(\boldsymbol{\beta}_{jk}; \boldsymbol{\tau}_{jk}, \boldsymbol{\alpha}_{jk})}{\partial \boldsymbol{\beta}_{jk} \partial \boldsymbol{\beta}_{jk}^T} \end{aligned}$$

Inference for parameters $\boldsymbol{\beta}_{jk}$ can, under suitable regularity conditions [Walker, 1969], be based on the asymptotic normality of the posterior distribution

$$\boldsymbol{\beta}_{jk} | \mathbf{y} \stackrel{a}{\sim} \mathcal{N}(\hat{\boldsymbol{\beta}}_{jk}, \mathbf{H}(\hat{\boldsymbol{\beta}}_{jk})^{-1}) \quad (5.4)$$

with $\hat{\boldsymbol{\beta}}_{jk}$ being the posterior mode estimate and $\mathbf{H}(\hat{\boldsymbol{\beta}}_{jk})$ the hessian matrix. The marginal asymptotic normality of each BAMLSS parameter estimator is crucial for applying BiCAM at the second step.

5.2.2 Second step: Bivariate CAM

In the first step, we have fitted BAMLSS models $Y = f_y(X) + g_y(X)\varepsilon_y$ and $X = f_x(Y) + g_x(Y)\varepsilon_x$ from which we get estimates $\hat{\boldsymbol{\beta}}_{Y|X}$ and $\hat{\boldsymbol{\beta}}_{X|Y}$ respectively. Denote $\tilde{\boldsymbol{\beta}}_{Y|X}$ (respect. $\tilde{\boldsymbol{\beta}}_{X|Y}$) the corresponding fitted vector that is composed by three sets of fitted values $\tilde{\mu}_{Y|X}$ (respect. $\tilde{\mu}_{X|Y}$) for the location parameter, $\tilde{\sigma}_{Y|X}$ (respect. $\tilde{\sigma}_{X|Y}$) for the scale parameter, $\tilde{\xi}_{Y|X}$ (respect. $\tilde{\xi}_{X|Y}$) for the shape parameter, each of size n . Suppose n large, from (5.4), each fitted values set

can be supposed to come from a normal distribution, so that we can use Bi-CAM, an appropriate and efficient approach to infer causality in the Gaussian context. The basic idea of the causal additive models (CAM) is to learn the causal direction from an observational joint distribution by assuming that the effect can be written as some function of the cause plus additive noise, which is independent from the cause. In what follows we detail the second-step process for the location parameter μ but we also process in a similar way for the scale and shape parameters. We aim at testing the causal models “ X causes $\tilde{\mu}_{Y|X}$ ” ($X \rightarrow \tilde{\mu}_{Y|X}$) and “ Y causes $\tilde{\mu}_{X|Y}$ ” ($Y \rightarrow \tilde{\mu}_{X|Y}$) by regressing $\hat{\mu}_{Y|X}$ on X and respectively $\hat{\mu}_{X|Y}$ on Y that is,

$$\tilde{\mu}_{Y|X} = v_{Y|X}(X) + \varepsilon_{Y|X} \quad (5.5)$$

$$\tilde{\mu}_{X|Y} = v_{X|Y}(Y) + \varepsilon_{X|Y} \quad (5.6)$$

where

- $v_{Y|X}(\cdot)$ and $v_{X|Y}(\cdot)$ are smooth functions $\mathbb{R} \rightarrow \mathbb{R}$
- $E(v_{Y|X}(X)) = 0$; $E(v_{X|Y}(Y)) = 0$
- $\varepsilon_{Y|X} \sim \mathcal{N}(0, \sigma_{Y|X}^2)$ with $\sigma_{Y|X}^2 > 0$ and $\varepsilon_{X|Y} \sim \mathcal{N}(0, \sigma_{X|Y}^2)$ with $\sigma_{X|Y}^2 > 0$
- $\varepsilon_{Y|X}, \varepsilon_{X|Y}$ independent.

If $v_{Y|X}(\cdot) \neq 0 \Rightarrow \exists$ a causal link between X and $\tilde{\mu}_{Y|X}$ and if $v_{X|Y}(\cdot) \neq 0 \Rightarrow \exists$ a causal link between Y and $\tilde{\mu}_{X|Y}$. It may happen that both directions are significant, or in other words both $\hat{v}_{X|Y}$ and $\hat{v}_{Y|X}$ are significantly different from 0. To decide the direction, we calculate an independence score between the residuals $r_{Y|X} = \hat{v}(\tilde{\mu}_{Y|X}) - X$ and the regressor X for model (5.5), and between $r_{X|Y} = \hat{v}(\tilde{\mu}_{X|Y}) - Y$ and Y for model (5.6), and choose the model with the highest likelihood independence score. This measure of independence relies on the fact that if $X \rightarrow \tilde{\mu}_{Y|X}$ then $\tilde{\mu}_{Y|X}|X \perp\!\!\!\perp X$ therefore $r_{Y|X} \perp\!\!\!\perp X$. The so-estimated causal effect is consistent [Peters et al., 2014]. If $v_{Y|X}(\cdot)$ is non-linear then $v_{Y|X}(\cdot)$ is identifiable from the joint distribution of $\tilde{\mu}_{Y|X}$ and X that is, the causal structure of variables can be uniquely estimated using the observational data. Same for $v_{X|Y}(\cdot)$. We apply a similar process to the scale and shape parameters. The final direction is $X \rightarrow Y$ if the sum of the independence scores for $X \rightarrow \tilde{\mu}_{Y|X}$, $X \rightarrow \tilde{\sigma}_{Y|X}$ and $X \rightarrow \tilde{\xi}_{Y|X}$ is higher than the sum of the independence scores for $Y \rightarrow \tilde{\mu}_{X|Y}$, $Y \rightarrow \tilde{\sigma}_{X|Y}$ and $Y \rightarrow \tilde{\xi}_{X|Y}$. The underlying assumption our process rely on is that if X causes Y , then X has an effect on the characteristics of the distribution of Y , that is on its location, scale and shape parameters.

5.3 Simulation study

To assess our method and compare it to competitive approaches, we ran an extensive simulation study. We suppose that X is the cause of Y and that the causal relation is explained by a multiplicative noise model:

$$Y = f(X) + g(X)\varepsilon \quad (5.7)$$

Relying on BAMLSS, our method can handle any response distribution. In the simulation study we consider three different distributions for X namely, the normal distribution ($X \sim \mathcal{N}(0, 1)$), the log-normal distribution ($X \sim \text{Log-}\mathcal{N}(1, 0.25)$) and the generalized pareto distribution ($X \sim \text{GPD}(0.2, 5, 0.5)$) which happens to be useful in finance in a risk management perspective. For the choice of $f(\cdot)$ and $g(\cdot)$ we consider the following functions: $-x^2$, $3x^3 + x$, $\tanh(x)$, $\log(|x|)$ and $\sqrt{|x|}$. We generate 200 samples of $y_i, i = 1, \dots, n$ from (5.7) with the different functions $f(\cdot)$ and $g(\cdot)$, and four different levels of standard deviation of the noise term ($\varepsilon \sim \mathcal{N}(0, sd)$), each sample of size $n = 500$. We apply our CHM method on each of the 200 simulated data to find the causality direction and compare its performance with four alternative methods which are

1. CAM [Bühlmann et al., 2014]: non-linear-Gaussian structural equation models.
2. Linear non-Gaussian acyclic causal model (LinGAM) [Shimizu et al., 2006]: relies on independent component analysis with the additional assumption that disturbance variables have non-Gaussian distributions of non-zero variances (implementation of Peters et al. [2014]).
3. Regression with subsequent independence test (RESIT) [Peters et al., 2014]: is based on the fact that for each node X , the corresponding noise variable ε_x is independent of all non-descendants of X .
4. Quantile-Based Causal Discovery (QCCD) [Tagasovska et al., 2018]: is based on the link between Kolmogorov complexity and quantile scoring using a nonparametric conditional quantile estimator based on copulas, hence placing no restrictive assumptions about the joint distribution. The QCCD method can handle any distribution including heteroscedastic distributions.

These competitive approaches are the most recent and commonly recognized methods in causality. CAM, LinGAM and RESIT are additive noise models therefore can hardly handle heteroscedasticity. The QCCD method is the only one allowing a fair comparison since it deals with heteroscedasticity.

Tables 5.1, 5.2 and 5.3 provide a summary of accuracy of the compared methods based on simulating samples from (5.7) with the different functions $f(\cdot)$ and $g(\cdot)$ and with X coming from the normal distribution (Table 5.1), the log-normal distribution (Table 5.2) and the GPD (Table 5.3). The accuracy is the percentage of times the method discovers the true causality direction. Each figure of these tables is a mean of four accuracies corresponding to four different levels of standard deviation of the noise term ε . Globally our approach performs much better than the competitive approaches in the log-normal and GPD cases since these cases have more heteroscedasticity. In case of the normal distribution we can distinguish two different scenarios according to the intensity of the multiplicativity in $f(x) + g(x)\varepsilon$. We find that whenever $g(x)$ dominates $f(x)$ that is $f(x) \ll g(x)$ (see examples in Figure 5.3 left panel) the multiplicative part dominates and CHM performs better. This is the heteroscedasticity case we typically find in finance and want to handle. Figure

5.1 shows boxplots of accuracy for the five compared methods for the normal distribution with a special case where $g(x)$ dominates and for different values of the standard deviation of the noise ε . In this highly multiplicative case, CHM gets a mean accuracy of 0.94. It slightly improves when the noise standard deviation increases, re-enforcing the multiplicative part. In the case where $f(x) \gg g(x)$, where $g(x)$ is negligible compared to $f(x)$ (see examples in Figure 5.3 right panel) we tend to an additive noise model, hence the performance of CHM is affected but still good. Figure 5.2 represents the boxplots of accuracy of the five methods in the normal case with $f(x)$ dominating and for different values of the standard deviation of the noise ε . Again, the CHM accuracy improves when the noise standard deviation increases. The average accuracy for CHM is 0.76 but LinGAM and RESIT perform better reaching both the accuracy of 1 on average. In GPD and log-normal scenarios CHM performs well and better than all compared methods in most of the cases with the accuracy reaching 0.97 on average.

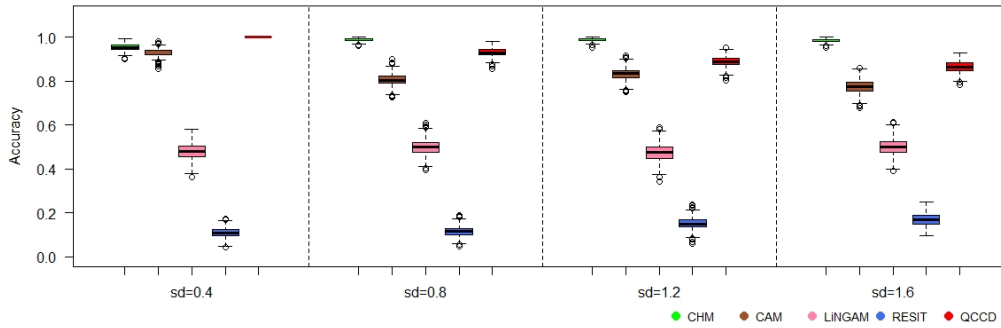


Figure 5.1: Boxplots of estimated accuracy of different methods for the model $Y = -X^2 + (3X + X)\varepsilon$ on 200 samples of size $n = 500$, where $X \sim \mathcal{N}(0, 1)$, and $\varepsilon \sim \mathcal{N}(0, sd)$

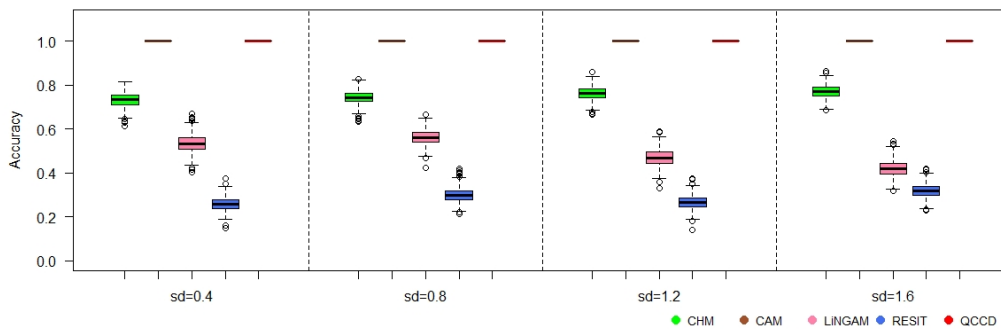


Figure 5.2: Boxplots of estimated accuracy of different methods for the model $Y = -X^2 + \sqrt{|X|}\varepsilon$ on 200 samples of size $n = 500$, where $X \sim \mathcal{N}(0, 1)$, and $\varepsilon \sim \mathcal{N}(0, sd)$

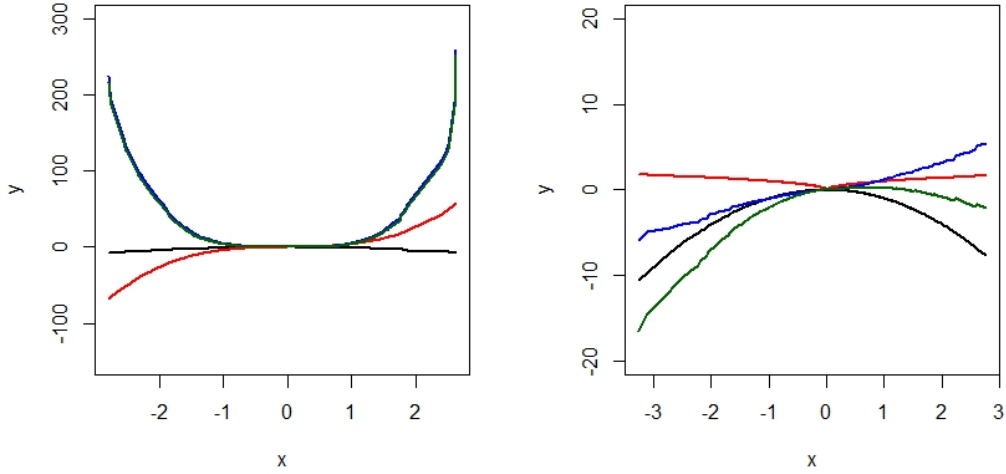


Figure 5.3: $X \sim \mathcal{N}(0, 1)$ and $\varepsilon \sim \mathcal{N}(0, 1.2)$; We consider different relations $Y = f(X) + g(X)\varepsilon$ with multiplicative dominant constructions (left panel) that are $y = -x^2$, $y = x^3 + x$, $y = (x^3 + x)\varepsilon$, $y = -x^2 + (x^3 + x)\varepsilon$ and additional dominant constructions (right panel) that are $y = -x^2$, $y = \sqrt{|x|}$, $y = \sqrt{|x|}\varepsilon$, $y = -x^2 + \sqrt{|x|}\varepsilon$

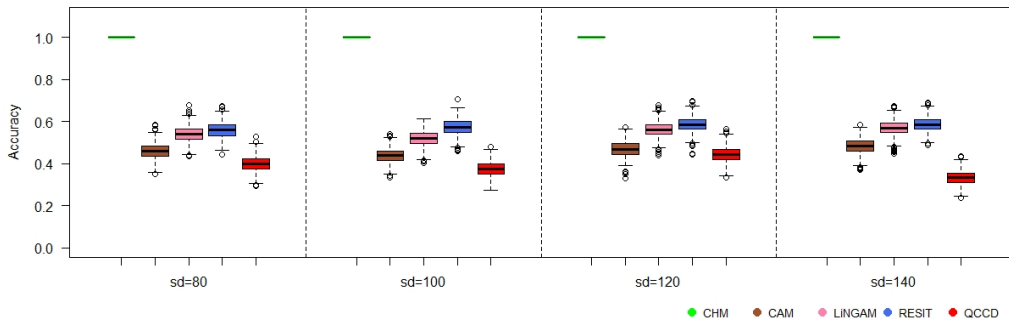


Figure 5.4: Boxplots of estimated accuracy of different methods for the model $Y = \tanh(X) + \sqrt{|X|}\varepsilon$ on 200 samples of size $n = 500$, where $X \sim GPD(0.2, 5, 0.5)$, and $\varepsilon \sim \mathcal{N}(0, sd)$

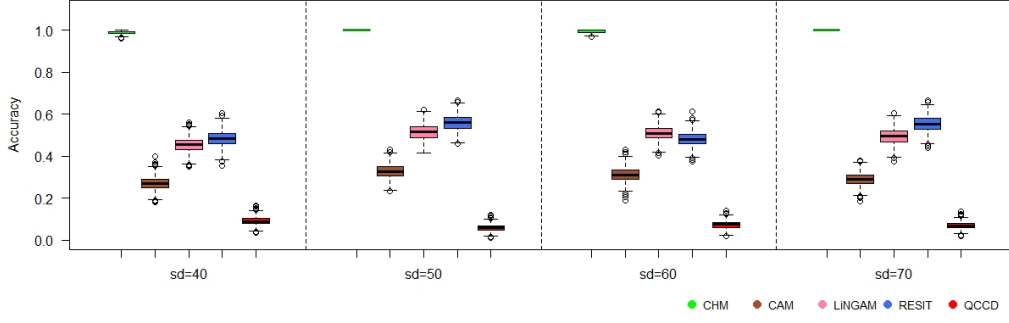


Figure 5.5: Boxplots of estimated accuracy of different methods for the model $Y = \log(|X|) + \sqrt{(|X|)}\varepsilon$ on 200 samples of size $n = 500$, where $X \sim \text{Log-}\mathcal{N}(1, 0.25)$, and $\varepsilon \sim \mathcal{N}(0, sd)$

	CHM	CAM	LinGAM	RESIT	QCCD
$Y = -X^2 + (3X^3 + X)\varepsilon$	0.94	0.90	0.51	0.13	0.98
$Y = -X^2 + \tanh(X)\varepsilon$	0.68	1	0.51	0.29	1
$Y = -X^2 + \log(X)\varepsilon$	0.61	1	0.54	0.23	1
$Y = -X^2 + \sqrt{ X }\varepsilon$	0.76	1	0.53	0.30	1
$Y = 3X^3 + X - X^2\varepsilon$	0.76	1	0.53	0.30	1
$Y = 3X^3 + X + \tanh(X)\varepsilon$	0.96	1	1	0.03	0.29
$Y = 3X^3 + X + \log(X)\varepsilon$	0.93	1	1	0.01	1
$Y = 3X^3 + X + \sqrt{ X }\varepsilon$	0.95	1	1	0.02	0.24
$Y = \tanh(X) - X^2\varepsilon$	0.95	1	1	0.02	0.24
$Y = \tanh(X) + (3X^3 + X)\varepsilon$	0.97	0.58	0.53	0.70	0.84
$Y = \tanh(X) + \log(X)\varepsilon$	0.55	0.13	0.62	0.19	0.14
$Y = \tanh(X) + \sqrt{ X }\varepsilon$	0.21	0.99	0.02	0	0.96
$Y = \log(X) - X^2\varepsilon$	0.21	0.99	0.02	0	0.96
$Y = \log(X) + (3X^3 + X)\varepsilon$	0.94	0.91	0.50	0.15	1
$Y = \log(X) + \tanh(X)\varepsilon$	0.36	1	0.48	0.40	1
$Y = \log(X) + \sqrt{ X }\varepsilon$	0.46	1	0.49	0.37	1
$Y = \sqrt{ X } - X^2\varepsilon$	0.47	1	0.49	0.37	1
$Y = \sqrt{ X } + (3X^3 + X)\varepsilon$	0.47	1	0.49	0.37	1
$Y = \sqrt{ X } + \tanh(X)\varepsilon$	0.46	1	0.48	0.37	1
$Y = \sqrt{ X } + \log(X)\varepsilon$	0.25	1	0.49	0.32	1

Table 5.1: Accuracy on gaussian simulated data

	CHM	CAM	LinGAM	RESIT	QCCD
$Y = -X^2 + (3X^3 + X)\varepsilon$	0.98	0.03	0.88	0.04	0
$Y = -X^2 + \tanh(X)\varepsilon$	0.95	0.13	1	0.98	0.68
$Y = -X^2 + \log(X)\varepsilon$	0.94	0.01	1	0.16	0
$Y = -X^2 + \sqrt{ X }\varepsilon$	0.97	0.04	0.95	0.69	0.09
$Y = 3X^3 + X - X^2\varepsilon$	0.94	0.30	0.99	0.01	0
$Y = 3X^3 + X + \tanh(X)\varepsilon$	0.91	0.72	1	0.79	0.76
$Y = 3X^3 + X + \log(X)\varepsilon$	0.90	0.43	1	0.05	0.01
$Y = 3X^3 + X + \sqrt{ X }\varepsilon$	0.92	0.14	1	0.77	0.16
$Y = \tanh(X) - X^2\varepsilon$	1	0	0.51	0.37	0
$Y = \tanh(X) + (3X^3 + X)\varepsilon$	1	0.01	0.55	0.35	0
$Y = \tanh(X) + \log(X)\varepsilon$	1	0.01	0.52	0.50	0
$Y = \tanh(X) + \sqrt{ X }\varepsilon$	1	0.33	0.49	0.52	0.06
$Y = \log(X) - X^2\varepsilon$	1	0	0.53	0.40	0
$Y = \log(X) + (3X^3 + X)\varepsilon$	1	0.02	0.48	0.33	0
$Y = \log(X) + \tanh(X)\varepsilon$	1	0.52	0.53	0.52	0.49
$Y = \log(X) + \sqrt{ X }\varepsilon$	1	0.30	0.49	0.52	0.07
$Y = \sqrt{ X } - X^2\varepsilon$	1	0	0.51	0.41	0
$Y = \sqrt{ X } + (3X^3 + X)\varepsilon$	1	0.06	0.53	0.37	0.05
$Y = \sqrt{ X } + \tanh(X)\varepsilon$	1	0.48	0.49	0.49	0.49
$Y = \sqrt{ X } + \log(X)\varepsilon$	0.99	0	0.02	0.02	0

Table 5.2: Accuracy on Log-normal simulated data

	CHM	CAM	LinGAM	RESIT	QCCD
$Y = -X^2 + (3X^3 + X)\varepsilon$	1	0.11	0.53	0.87	0.01
$Y = -X^2 + \tanh(X)\varepsilon$	0.78	0.06	1	0.99	0.88
$Y = -X^2 + \log(X)\varepsilon$	0.85	0.04	1	0.99	0.58
$Y = -X^2 + \sqrt{ X }\varepsilon$	0.89	0.07	0.94	0.92	0.60
$Y = 3X^3 + X - X^2\varepsilon$	0.95	0.40	0.98	0.88	0.05
$Y = 3X^3 + X + \tanh(X)\varepsilon$	0.94	0.91	1	0.59	0.85
$Y = 3X^3 + X + \log(X)\varepsilon$	0.95	0.64	1	0.79	0.69
$Y = 3X^3 + X + \sqrt{ X }\varepsilon$	0.95	0.43	1	0.90	0.74
$Y = \tanh(X) - X^2\varepsilon$	1	0.02	0.48	0.86	0
$Y = \tanh(X) + (3X^3 + X)\varepsilon$	1	0.09	0.51	0.86	0.01
$Y = \tanh(X) + \log(X)\varepsilon$	1	0.45	0.53	0.55	0.34
$Y = \tanh(X) + \sqrt{ X }\varepsilon$	1	0.46	0.55	0.58	0.39
$Y = \log(X) - X^2\varepsilon$	1	0.02	0.49	0.88	0.01
$Y = \log(X) + (3X^3 + X)\varepsilon$	1	0.10	0.52	0.87	0.01
$Y = \log(X) + \tanh(X)\varepsilon$	0.99	0.52	0.52	0.53	0.51
$Y = \log(X) + \sqrt{ X }\varepsilon$	0.99	0.44	0.52	0.54	0.39
$Y = \sqrt{ X } - X^2\varepsilon$	1	0.02	0.52	0.87	0
$Y = \sqrt{ X } + (3X^3 + X)\varepsilon$	1	0.12	0.54	0.83	0.01
$Y = \sqrt{ X } + \tanh(X)\varepsilon$	0.99	0.51	0.52	0.52	0.51
$Y = \sqrt{ X } + \log(X)\varepsilon$	1	0.45	0.54	0.56	0.35

Table 5.3: Accuracy on GPD simulated data

5.4 Stock market indices

In this section, our method is used to determine the intrinsic casual mechanism between shares and indices. An index being composed of weighted stock prices, the share is a “parent” and the index, the “child”. We expect our method to retrieve the “share \rightarrow index” causal direction. The analysis is carried out on the **S&P 500**, **CAC 40** and **Nikkei** indices and, for each index, on five of its shares with important weights. The indices and shares are listed in the first column of Table 5.5. The data consists of daily log-returns from 2008 to 2018.

5.4.1 Pairwise exploration

We apply our CHM method to discover a causal direction for each “share-index” pair. Figure 5.6 illustrates three examples of these pairs. The scatter plots of the other pairs appear in Appendix. To test the heteroscedaticity of all pairs we use the Breusch-Pagan test [Breusch and Pagan, 1979; Koener, 1981], the results are reported in Table 5.4. The considered data is globally heteroscedastic, except for Nikkei where the test based on linear regression could not capture the singularity of the joint distribution shaped like a cross (Figure 5.6). Table 5.5 displays the estimated causal direction between the pairs of share and index for our method and for the four other methods. Using our methodology, in almost all cases we find the correct causal effect of the share on the index. In the Nikkei case, apart from CHM and QCCD, all the other methods fail to retrieve the correct direction. This can be explained by the unusual and complex structure between Nikon and Nikkei as shown in the right panel of Figure 5.6. Indeed, the bivariate structure for this pair is much less elliptical than for other pairs like the one for Coca-Cola and S&P500 (left panel of Figure 5.6) and that shows more evidence for bivariate Gaussian distribution. Both our method and QCCD capture these complex links by exploring the causal mechanism at different moments or quantiles of the distribution.

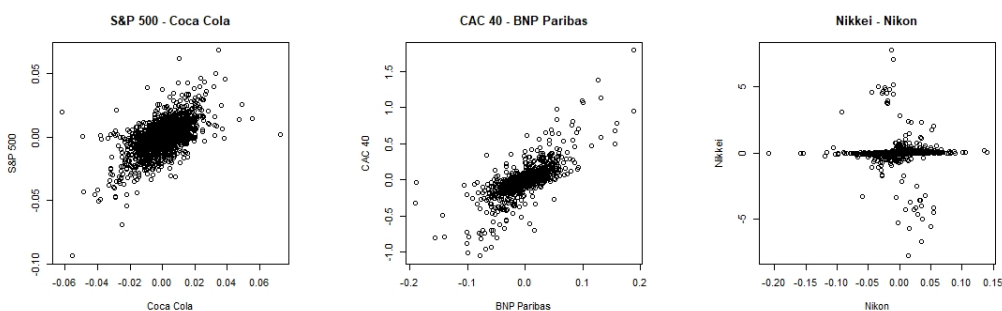


Figure 5.6: Scatter plot of indices and stocks

5.4.2 Extension to multivariate case

The previous section was about discovering pairwise underlying causal structure between one share and the index. In this section, we propose an extension of our pairwise CHM to the multivariate case and explore the causal structure between different shares. This extension leads to graphical modelling that consists of fitting graph that satisfies a set of conditional independence relations,

Pairs	Heteroscedasticity	p-value
S&P 500 ~ Cola cola	Yes	1.60 E-7
S&P 500 ~ Medtronic	Yes	4.87 E-4
S&P 500 ~ Walt Disney Company	Yes	2.13 E-3
S&P 500 ~ Caterpillar	Yes	0.023
S&P 500 ~ Bank of America	Yes	1.26 E-6
CAC 40 ~ BNP Paribas	Yes	0.087
CAC 40 ~ Legrand	Yes	<2.2 E-16
CAC 40 ~ Orange	No	0.68
CAC 40 ~ Sanofi	Yes	0.035
CAC 40 ~ Total	Yes	2.18 E-4
Nikkei ~ Nikon	No	0.49
Nikkei ~ Inpex	Yes	2.25 E-3
Nikkei ~ Japan Tobacco	No	0.96
Nikkei ~ Tosoh	No	0.61
Nikkei ~ Toyota	No	0.11

Table 5.4: Results of heteroscedastic test applied on stock market indices data

Pairs	CHM	CAM	LinGAM	RESIT	QCCD
S&P 500 - Cola cola	←	←	→	→	←
S&P 500 - Medtronic	←	←	→	→	←
S&P 500 - Walt Disney Company	←	←	→	→	←
S&P 500 - Caterpillar	←	←	→	→	←
S&P 500 - Bank of America	→	←	→	←	←
CAC 40 - BNP Paribas	←	→	←	→	→
CAC 40 - Legrand	→	→	←	←	←
CAC 40 - Orange	←	→	←	→	→
CAC 40 - Sanofi	←	←	→	→	→
CAC 40 - Total	←	→	→	→	→
Nikkei - Nikon	←	→	→	→	←
Nikkei - Inpex	←	→	→	→	←
Nikkei - Japan Tobacco	←	→	→	→	←
Nikkei - Tosoh	←	→	→	←	←
Nikkei - Toyota	←	→	→	←	←

Table 5.5: Estimated causal direction on stock market indices data

known as markov property [Maathuis et al., 2009]. We use the idea suggested by Goudet et al. [2018] which is that pairwise and CPDAG (the skeleton and the v-structures of a graphical model) learning procedures are complementary in the multivariate analysis. To get the edges, for each pair of variables X and Y , we remove the effect of all remained covariates using gam models [Hastie and Tibshirani, 1990; Wood, 2017] and fit our CHM method on the obtained pair of gam residuals r_x and r_y to get the edge between X and Y . Then we rank the resulting edges according to the standardized score defined as $score = \frac{|s_{y|x} - s_{x|y}|}{\max(s_{y|x}, s_{x|y})}$, where $s_{y|x}$ and $s_{x|y}$ are the scores obtained by fitting

CHM models $r_y = f_y(r_x) + g_y(r_x)\varepsilon_y$ and $r_x = f_x(r_y) + g_x(r_y)\varepsilon_x$ respectively. Finally we include the ranked edges in the graph sequentially, starting from the highest score while checking the acyclicity of the resulting graph after each addition.

The so obtained directed acyclic graph (DAG) corresponding to S&P 500 shares is illustrated in Figure 5.7 for the normal case, meaning using the observed log returns. The normal DAGs of CAC 40 and Nikkei shares appear in Appendix. The darker and thicker the arrow, the stronger the causal effect. We note in fig. 5.7 that the Bank of America and Caterpillar are mainly “parent” whereas Coca Cola and Medtronic are essentially “Child”.

After exploring the normal situation, we were interested in analysing the extremes, especially negative log returns extremes representing small financial crisis. We use “Peak Over Threshold” method [Coles et al., 2001], which consists of selecting values exceeding a certain threshold, namely the 0.85 quantile of negative log returns and fit a GPD. Figure 5.8 provides the DAG obtained after applying the multivariate CHM to the extremes of negative log returns of shares of S&P 500. Unlike the normal case, the DAG of extremes shows that Bank of America, Caterpillar and Disney are essentially “Child” meaning effects, while Coca Cola and Medtronic have become “parent” meaning causes. The DAGs of CAC 40 and Nikkei shares extreme negative log returns appear in Appendix. These causal graphs are interesting since they provide the causal structures in different scenarios, and could be useful for asset and financial portfolio management.

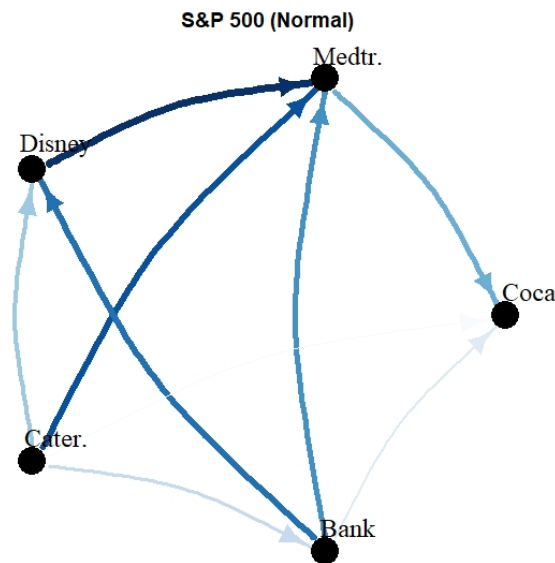


Figure 5.7: DAGs for S&P 500 shares in the normal case

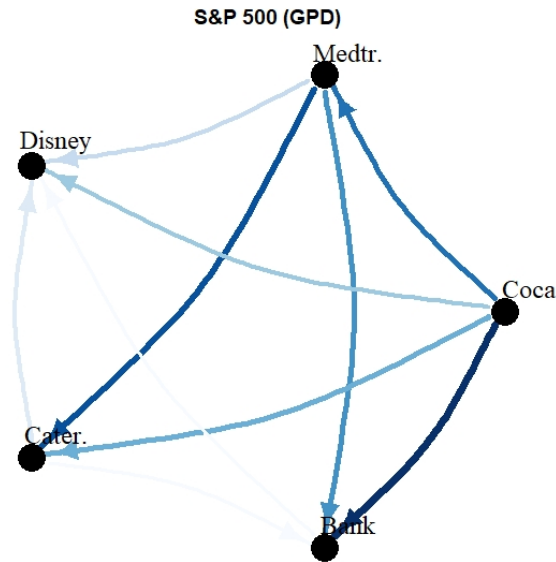


Figure 5.8: DAGs for S&P 500 shares in extreme case

5.5 Conclusion

In this paper we provide a two-steps method CHM based on BAMLSS and CAM, that allows the causal inference of heteroscedastic data through causal multiplicative noise models and can handle any response distribution. We assess our method on both simulated and real financial data. In the simulation study we consider normal, generalized pareto and lognormal distributions. In the highly multiplicative of the normal case CHM has a mean accuracy of 0.94, and in the GPD and lognormal scenarios CHM reaches the accuracy of 0.97 on average. The application of CHM on heteroscedastic financial data is motivated by the fact that our method is not conditioned on time and investigate the causal relation not only in mean but in higher moments such as the scale and the shape. The analyse of financial indices data we find an un-lagged causal effect of the shares on the index they compose. Hence, CHM method is useful for inferring causality for any response distribution.

5.6 Appendix

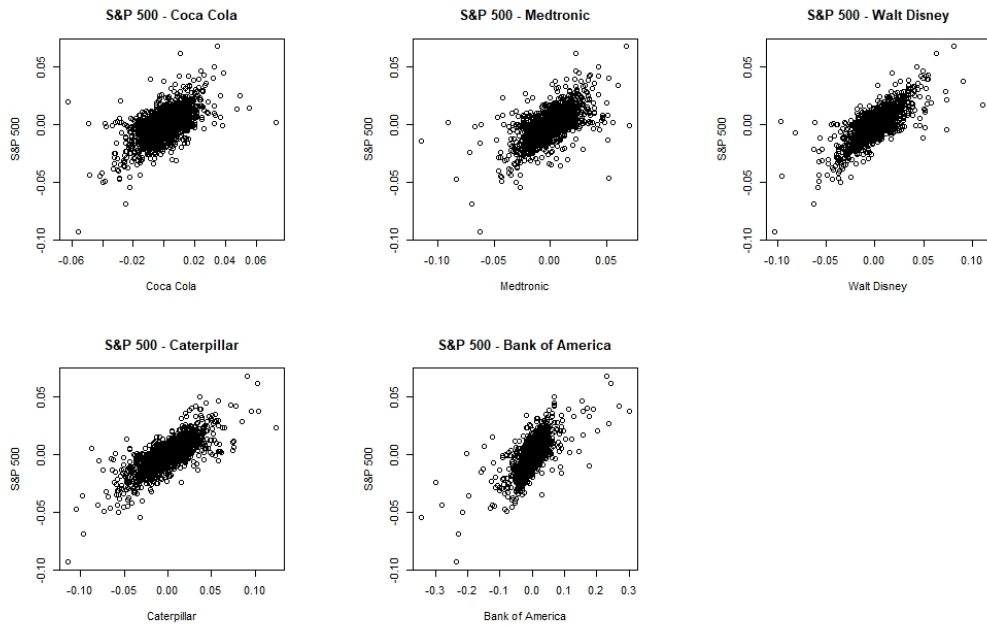


Figure 5.9: Scatter plot of S&P 500 and stocks log returns

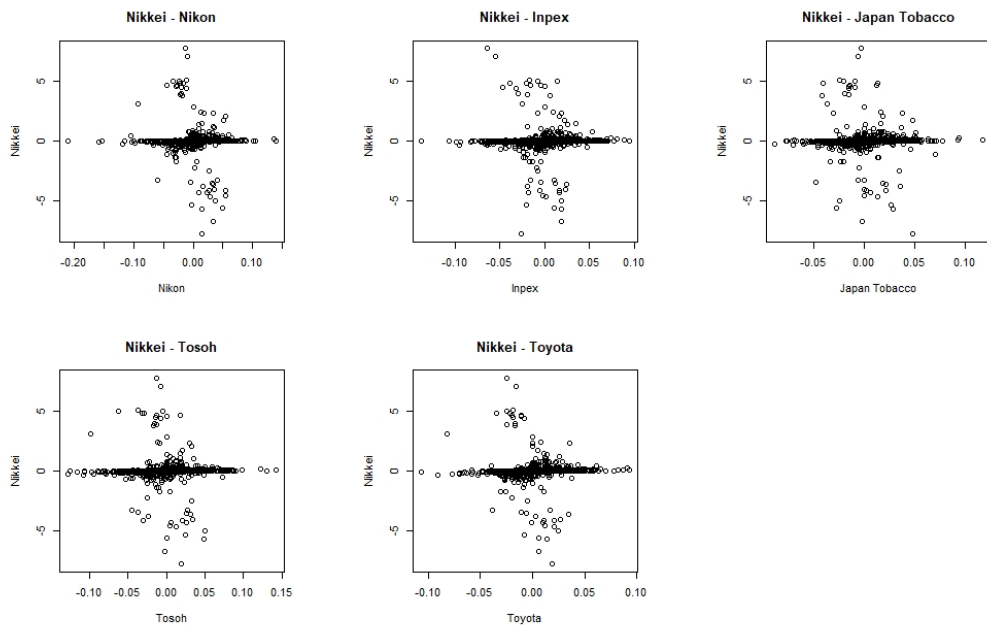


Figure 5.11: Scatter plot of Nikkei and stocks log returns

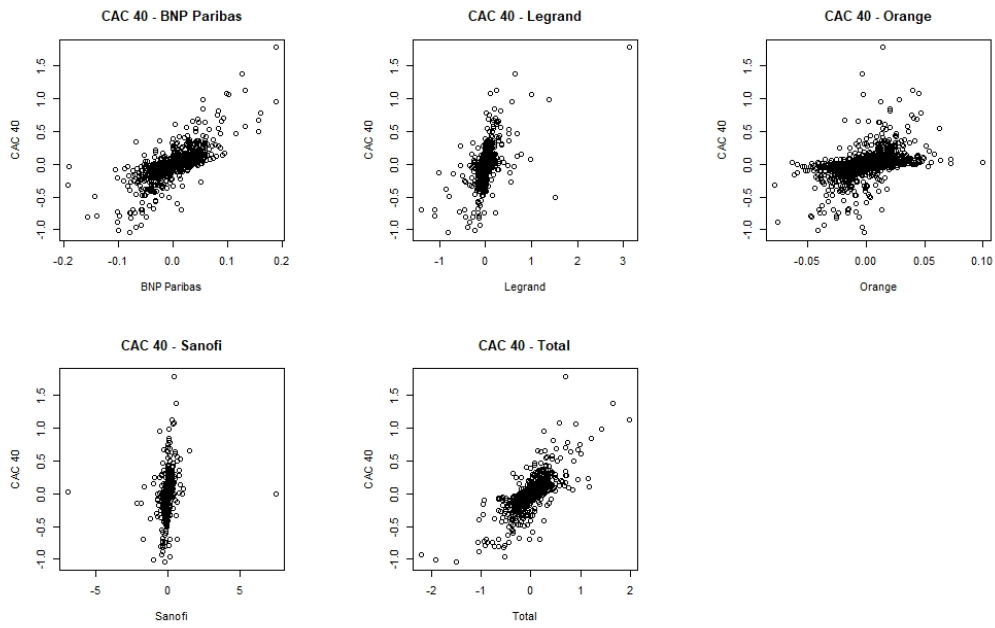


Figure 5.10: Scatter plot of CAC 40 and stocks log returns

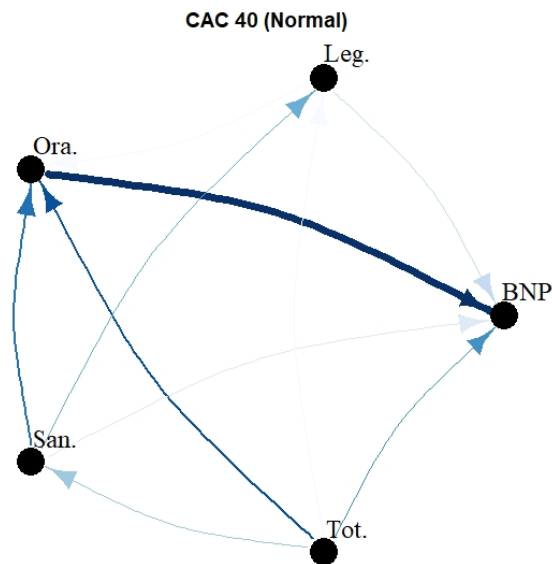


Figure 5.12: DAGs for CAC 40 shares in the normal case

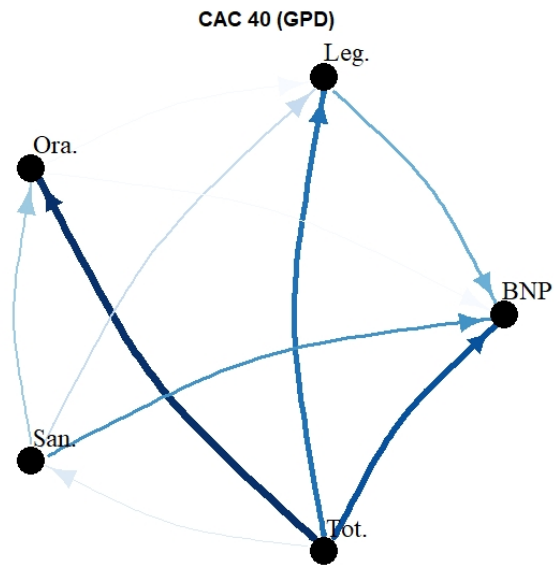


Figure 5.13: DAGs for CAC 40 shares in extreme case

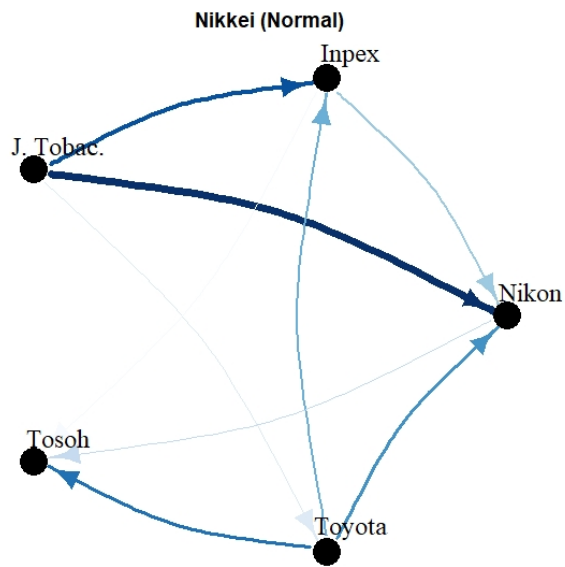


Figure 5.14: DAGs for Nikkei shares in the normal case

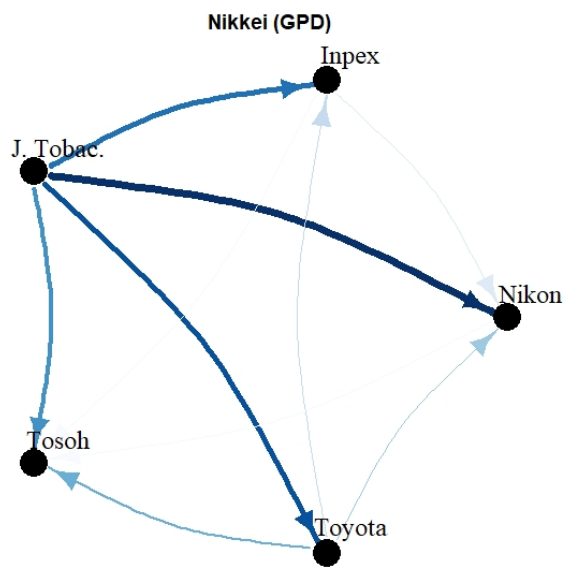


Figure 5.15: DAGs for Nikkei shares in extreme case

Bibliography

- R. Adhikari and R. K Agrawal. Forecasting strong seasonal time series with artificial neural networks. *Journal of Scientific Industrial Research*, 71(Oct): 657–666, 2012.
- L. Anselin. The future of spatial analysis in the social sciences. *Geographic information sciences*, 5(2):67–76, 1999.
- M. Ashraf, J. C. Loftis, and K. G. Hubbard. Application of geostatistics to evaluate partial weather station networks. *Agricultural and forest meteorology*, 84(3-4):255–271, 1997.
- M. D. Atchison, K. C. Butler, and R. R. Simonds. Nonsynchronous security trading and market index autocorrelation. *The Journal of Finance*, 42(1): 111–118, 1987.
- J. B. Ayers. *Handbook of supply chain management*. Auerbach publications, 2006.
- J. Blackhurst, K. S. Dunn, and C. W. Craighead. An empirically derived framework of global supply resiliency. *Journal of business logistics*, 32(4): 374–391, 2011.
- T. S. Breusch and A. R Pagan. A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, pages 1287–1294, 1979.
- P. Bühlmann, J. Peters, and J. Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.
- F. Y. Chen and C. A. Yano. Improving supply chain performance and managing risk under weather-related demand uncertainty. *Management Science*, 56(8):1380–1397, 2010.
- J. Chen and L. Xu. Coordination of the supply chain of seasonal products. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 31(6):524–532, 2001.
- C. C. Chuang, C. M. Kuan, and H. Y. Lin. Causality in quantiles and dynamic stock return–volume relations. *Journal of Banking Finance*, 33(7):1351–1360, 2009.
- S. Coles, J. Bawa, L. Trenner, and P. Dorazio. *An introduction to statistical modeling of extreme values*. Springer, 2001.
- F. F. Costa. Big data in biomedicine. *Drug discovery today*, 19(4):433–440, 2014.
- R. Dahistrom, K. M. McNeilly, and T. W. Speh. Buyer-seller relationships in the procurement of logistical services. *Journal of the Academy of Marketing Science*, 24(2):110–124, 1996.

- A. A. Dick. Demand estimation and consumer welfare in the banking industry. *Journal of Banking Finance*, 32(8):1661–1676, 2008.
- D. Galles and J. Pearl. *Testing identifiability of causal effects*. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1995.
- D. Galles and J. Pearl. Axioms of causal relevance. *Artificial Intelligence*, 97(1-2):9–43, 1997.
- M. Gan, Y. Cheng, K. Liu, and G. L. Zhang. Seasonal and trend time series forecasting based on a quasi-linear autoregressive model. *Applied Soft Computing*, 24:13–18, 2014.
- A. Gelman, G. O. Roberts, and W. R. Gilks. Efficient metropolis jumping rules. *Bayesian statistics*, 5(599-608):42, 1996.
- O. Goudet, D. Kalainathan, P. Caillou, I. Guyon, D. Lopez-Paz, and M. Sebag. Learning functional causal models with generative neural networks. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, Springer, Cham:39–80, 2018.
- C. W. Granger. Some properties of time series data and their use in econometric model specification. *Journal of econometrics*, 16(1):121–130, 1981.
- C. W. Granger and O. Morgenstern. Spectral analysis of new york stock market prices 1. *Kyklos*, 16(1):1–27, 1963.
- T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, 1990.
- W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.
- J. Heikkilä. From supply to demand chain management: efficiency and customer satisfaction. *Journal of operations management*, 20(6):747–767, 2002.
- E. Hofmann and E. Rutschmann. Big data analytics and demand forecasting in supply chains: a conceptual analysis. *The International Journal of Logistics Management*, 29(2):739–766, 2018.
- P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, pages 689–696, 2009.
- A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*. John Wiley & Sons, 2004.
- R. Koenker. A note on studentizing a test for heteroscedasticity. *Journal of Econometrics*, 17(1):107–112, 1981.
- M. H. Maathuis, M. Kalisch, and P. Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

- F. H. Messerli. Chocolate consumption, cognitive function, and nobel laureates. *The New England Journal of Medicine*, 367(16), 2012.
- N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.
- M. Morris and C. R. Carter. Relationship marketing and supplier logistics performance: an extension of the key mediating variables model. *Journal of Supply Chain Management*, 41(4):32–43, 2005.
- M. Moussaïd, D. Helbing, and G. Theraulaz. How simple rules determine pedestrian behavior and crowd disasters. *Proceedings of the National Academy of Sciences*, 108(17):6884–6888, 2011.
- R. M. Neal. Slice sampling. *Annals of statistics*, pages 705–741, 2003.
- C. Nowzohour and P. Bühlmann. Score-based causal learning in additive noise models. *Statistics*, 50(3):471–485, 2016.
- J. Pearl. *Causality*. Cambridge university press, 2009.
- J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.
- F. Provost and T. Fawcett. Data science and its relationship to big data and data-driven decision making. *Big data*, 1(1):51–59, 2013.
- R. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554, 2005.
- D. B. Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of educational Psychology*, 66(5):688, 1974.
- J. Sadeghi, S. M. Mousavi, and S. T. A. Niaki. Optimizing an inventory model with fuzzy demand, backordering, and discount using a hybrid imperialist competitive algorithm. *Applied Mathematical Modelling*, 40(15-16):7318–7335, 2016.
- S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.
- S. R. Singh and T. Kumar. Inventory optimization in efficient supply chain management. *International Journal of Computer Applications in Engineering Sciences*, 1, 2011.
- P. Spirtes and K. Zhang. Causal discovery and inference: concepts and recent methodological advances. *Applied informatics*, 3(1):3, 2016.
- D. M. Stasinopoulos and R. A. Rigby. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23(7):1–46, 2007.

- N. Tagasovska, T. Vatter, and V. Chavez-Demoulin. Nonparametric quantile-based causal discovery. *arXiv preprint*, arXiv:1801.10579, 2018.
- H. R. Thomas, D. R. Riley, and V. E. Sanvido. Loss of labor productivity due to delivery methods and weather. *Journal of construction engineering and management*, 125(1):39–46, 1999.
- K. J. Tierney. Business impacts of the northridge earthquake. *Journal of Contingencies and crisis management*, 5(2):87–97, 1997.
- N. Umlauf, N. Klein, and A. Zeileis. Bamlss: Bayesian additive models for location, scale and shape (and beyond). *Journal of Computational and Graphical Statistics*, just-accepted, 2017.
- W. Van der Aalst. *Process mining: data science in action*. Springer, Heidelberg, 2016.
- A. M. Walker. On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society, Series B (Methodological)*:80–88, 1969.
- S. N. Wood. *Generalized additive models: an introduction with R*. Chapman and Hall/CRC, 2017.
- K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 647–655, 2009.



Flora Babongo

Flora Babongo received a Bachelor’s degree in Economics from University of Lausanne in 2010 and a Master’s degree in Statistics from University of Neuchâtel in 2012. Before starting her PhD she worked two years for the Swiss Federal Office of Statistics. She is interested in applying statistics in different fields. She is currently pursuing her PhD studies in Information Systems at University of Lausanne, in the Faculty of Business and Economics. She has published several papers in international journals and conference proceedings.