

Forecasting migration trends to New Zealand

Dovile Tamosiunaite
STUDENT NUMBER: 2027087

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE SCHOOL OF
HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

Supervisor: dr. H. J. Brighton
Second Reader: dr. M. Alimardani

Tilburg University

School of Humanities and Digital Sciences Department of Cognitive Science &
Artificial Intelligence Tilburg, The Netherlands

May 2019

Abstract

One of the most complex and uncertain demographic processes to forecast is international migration flows. Focusing on the case of New Zealand, this thesis develops and analyzes predictive models using a new dataset that integrates migration flows and sentiment indicators. To construct this new dataset, sentiment data is extracted from the Global Database of Events, Language, and Tone (GDELT) and arrivals data - from the Statistical Department of New Zealand. Under the assumption that the news sentiment between the two countries can improve predictive accuracy, this thesis applies supervised machine learning models to predict monthly relative changes in arrivals to New Zealand. It is found that compared with the Naïve model the machine learning models can accurately predict at the country level and can generalize to some degree to unseen data when tested on the other country. Additionally, the models' performance is compared within countries in continents to find the underlying patterns. However, no single model has achieved an equally good result in terms of all experiments and the performance has shown to be dependent on a continent for which the predictions are made. As a result, the importance of sentiment in the best predictive models varies. It is found to have a minimal impact in Lasso and Ridge models while in Random Forest and partly Linear Regression the influence is greater.

Keywords: GDELT, machine learning, supervised learning, sentiment, migration, forecasting, sliding time window

Table of contents

Abstract.....	2
Introduction.....	5
1. RELATED WORK	8
1.1. SENTIMENT IN RESPECT OF MIGRATION	8
1.1.1. Negative aspects of migration	8
1.1.2. Factors affecting migration.....	8
1.2. MIGRATION PREDICTION	9
1.2.1. Predictive models.	9
1.2.2. Geolocated online data	11
2. EXPERIMENTAL SETUP	13
2.1. DESCRIPTION OF DATA.....	13
2.1.1. The Statistics Department of New Zealand database	13
2.1.2. The GDELT events database	14
2.1.3. Data pre-processing	16
2.1.4. Sliding time window approach	22
2.1.5. Additional features.	24
2.2. METHODOLOGY	25
2.2.1. Description of experiments.....	25
2.2.2. Models and tuning parameters.....	27
2.2.3. Evaluation metrics	30
3. RESULTS	32
3.1. RESEARCH QUESTION 1	32
3.2. RESEARCH QUESTION 2	37
3.2. RESEARCH QUESTION 3	39
4. DISCUSSION	43
4.1. REVIEW OF EXPERIMENTS.....	43
4.2. LIMITATIONS AND FUTURE WORK.....	45
5. CONCLUSION.....	48
References.....	49
Appendix A: List of countries.....	55
Appendix B: Missing months per year	56

Appendix C: Average tone before and after z-scores for Australia and the United States.....	57
Appendix D: Seasonality detection.....	58
Appendix E: Libraries.....	59
Appendix F: Results of experiment 1	60
Appendix G: Results of experiment 3.....	63

Introduction

One of the most complex and uncertain demographic processes to forecast is international migration flows. Migration is influenced by a range of different factors and due to that results are potentially inaccurate. Therefore, forecasts are biased as there is no perfect theory of migration itself. There are a few reasons why migration prediction is a challenging task. Forecasts are restricted by the availability and reliability of data, limited characteristics, length of prediction period and stability of trends (Bijak et al., 2015; Bohme, Groger, & Stohr, 2018; Zagheni, Garimella, Weber, & State, 2014). Usually, the data available on migration is unreliable and/or incomplete as it is not designed for that purpose (Buettner & Muenz, 2016; Disney et al., 2015). This is especially seen in developing countries where migration data is often unavailable due to the financial cost or lack of infrastructure related to collecting the data (Bohme, Groger, & Stohr, 2018; Wladyka, 2017). But even in developed countries if data on migration is gathered via longitudinal surveys there is an issue of a time lag to deliver statistical data. This leads to unreliable information for current analysis and policy-makers (Wladyka, 2017; Zagheni et al., 2014). In the presence of instability caused by a shortage of data series or events such as economic decline, political crises, ecological catastrophes, wars, there is no suitable model to accurately predict migration flows (Disney et al., 2015). Besides that, there is a huge diversity of unpredictable features which influences or prevents migration such as society positivity towards own economic success or anti-immigrant public sentiment (Buettner & Muenz, 2016; Lumpe, 2018; Gorinas & Pytlikova, 2015).

International migration is especially important in the demographic analysis as moving across borders is an endless global process. Immigrants are viewed as a solution that can reduce population decline and regenerate labor market (Gorinas & Pytlikova, 2015). However, for underdeveloped countries, the opposite effect is seen. The growing numbers of

departures cause a decline in population size and emigrants living and working in a foreign country are becoming a primary source of economic support (Buettner & Muenz, 2016). Therefore, the compound of the growing numbers of immigrants in the country and the public sentiment towards them has made assimilation of newcomers one of the primary concern regarding migration regulations (Gorinas & Pytlikova, 2015).

There have been many attempts to forecast migration (Azose & Raftery (2015); Raymer & Wisniowski (2018); Rojas-Romagosa & Bollen (2018), to name a few). Mostly a variety of statistical methods are used such as Autoregressive Integrated Moving Average (ARIMA), Vector Autoregression (VAR), Radiation and Gravity models. However, various machine learning models (Bayesian, Gradient Boosting Machine Regression, Lasso, Neural Networks) incorporated together with the online sources providing a real-time data such as Google Trends and the GDELT are used more widely too. Therefore, there is no universal model and methods to accurately predict migration are of huge importance to demographers, researchers, international organizations who analyze the complexities of migration and to local authorities to adjust their policies.

Following the need for building a better prediction model for migration, this thesis incorporates the news sentiment extracted from the GDELT and arrivals data extracted from the Statistical Department of New Zealand for the time period of 1979-2018. The choice for the aforementioned dataset is motivated by a high quality of data available on migration (a large time span, separate arrivals data, migration numbers for specific countries). The GDELT is chosen as it offers the largest near real-time information and provides a possibility to extract sentiment at the country-pair level. It is assumed that various events concerning changes in migration can be detected in the news articles. Thus, media serves as a source to learn about the public sentiment which might influence migration flow. The purpose of this

thesis is to compare different machine learning models' predictions against observed migration flow to New Zealand and to select the best approach for migration prediction.

This thesis seeks to address the following research questions:

1. How accurately do commonly used models predict migration flow to New Zealand and how can they be improved? Do predictive models generalize?
2. Are there any migration patterns within different continents in terms of migration flow to New Zealand? Is there a single best performing algorithm?
3. Is the sentiment an important determinant in migration prediction?

The remainder of this thesis is structured as follows. Section 1 consists of an in-depth literature review on sentiment and existing migration prediction methods in various researches. Section 2 provides a review of data sets used, pre-processing, detailed description of research questions and methodology. The results for the experiments are reported in Section 3. Section 4 provides a discussion of results and the recommendations for further improvements. Finally, conclusions and the main findings are summarized in Section 5.

1. RELATED WORK

1.1. SENTIMENT IN RESPECT OF MIGRATION

With an increase of migration numbers in the countries, the interest has grown in examining the public opinion towards immigrants. Especially towards analyzing the anti-immigrant sentiment followed by a discussion of imbalance between the public attitude and migration policies formation.

1.1.1. Negative aspects of migration. There are numerous studies that discuss the negative aspects of migration in relation to sentiment. Crush & Ramachandran (2010) have analyzed the problem of xenophobia phenomenon which causes the exclusion of minority class in everyday social life and nationally. This phenomenon in an extreme form is closely related to discrimination and violence towards immigrant groups. Researchers discuss that xenophobia is spreading with an increase of international immigrants in the countries of destination and that the national governments are failing to address this issue. Meseguer & Kemmerling (2018) have tested the anti-immigrant sentiment theories and have evaluated their relevance in developing countries. They have found evidence that the tax alterations are closely related to the negative sentiment toward immigrants. In addition to that, an increase of low skilled immigrants imposes strict immigration regulations. Sherkat & Lehman (2018) have investigated the sentiment toward immigrants in terms of religion in the United States. Their findings suggest a divide among Catholics based on ethnicity. White Catholics hold more negative views of immigration while non-white – more positive. Also, non-religious people are more positive toward immigrants. These views based on race and ethicality structures political commitments.

1.1.2. Factors affecting migration. Migration is being affected by many underlying features. Some of these for instance economic and political are widely reviewed in the literature (Disney et al., 2015). However, the public sentiment introduces another possible

explanation for migration fluctuations. Lumpe (2018) has analyzed public beliefs in social mobility and its impact on emigration intentions of highly skilled migrants in the main member countries of Organisation for Economic Co-operation and Development (OECD). The researcher concludes that the public beliefs regarding the movement of people could explain a particular country appeal for migrants. Thus, the positive public attitude towards own economic success attracts more highly skilled immigrants in the long term. Gorinas and Pytlikova (2015) have investigated if negative sentiment towards immigrants in a country affects the number of arrivals into OECD countries. Their results suggest that there is a negative attitude related to immigrants and there is a tendency of discrimination towards foreign workers when jobs are in a short supply. Immigrants with language proximity are more sensitive to a negative public attitude and media could serve as a source to learn about public opinion.

The public attitude towards immigrants especially anti-immigrant sentiment are extensively discussed in the literature and many types of research demonstrate sentiment influence on migration in various aspects. However, there is a lack of information on the role that media has on sentiment formation.

1.2. MIGRATION PREDICTION

Despite many attempts to predict international migration flows, there is no universal model and existing methods are relatively undeveloped. In a view of the fact that migration is relevant to demography, economy, policy changes, to name a few, it addresses the need for improving the forecasts.

1.2.1. Predictive models. Predominantly migration flows are forecasted by linking various methods, such as looking at historical data, using expert opinions and including other demographic and economic features (Bijak et al., 2015). This is mostly done in two ways: by using statistical, econometric or machine learning methods. The traditional econometric

methods include the ARIMA, an Autoregressive Distributed Lag (ADL), Radiation and Gravity models. Rojas-Romagosa and Bollen (2018) have used the structural gravity model to capture the Free Movement of People (FMP) in Europe. Researchers included country-specific determinants (unemployment rate, policy changes, economic growth) and country-pair elements (language, cultural and demographic distance, government regulations). Their findings suggest that FMP is one of the primary elements to explain migration within EU member states in 1960-2015 and highlights the importance of country-pair determinants. Bijak et al. (2019) compared a few different methods for migration prediction in the United Kingdom: ARIMA, ADL, methods incorporating historical errors. Some models have produced lower errors but no universal model has been found as specific method's result is dependent on the data being used. According to (Robinson & Dilkina, 2018) one of the drawbacks of these traditional models is an inability to capture a complex migration flow. Thus, they are outperformed by machine learning models. The researchers developed a few different machine learning models ("Extreme" Gradient Boosting Regression and Artificial Neural Network) and compared them to traditional Radiation and Gravity models to predict people movements globally and within states in the United States. Results show that machine learning models perform better than traditional ones. This was also found in the research done by (Azose & Raftery, 2015) who have projected net migration rates for all the countries by using Bayesian and Gravity methods in multiple experiments. The Bayesian model performance was better in all experiments. The aforementioned method is also used by (Raymer & Wisniowski, 2018) to measure migration flows in Sweden, South Korea, and Australia by incorporating age and gender-specific data. Authors recommend examining destination and origin countries by using distinct migration data. That is arrivals and departures instead of a net or gross migration.

1.2.2. Geolocated online data. The increasing availability of online metadata which includes information about geographical location offers an opportunity to track the spatial mobility of people and serves as an additional source regarding migration prediction. Zagheni et al. (2014) used social media's site "Twitter" data to track trends in migration. The author used geolocated information extracted from 500,000 users in OECD countries. The results indicate that the geolocated data can significantly enhance the knowledge about internal and international migration. Also, there is a potential to predict defining moments of change in migration patterns. However, as noted by (Bohme, Groger & Stohr, 2018) the contribution is primarily methodological to a degree that particular observations are used which cannot generalize, nor can be replicated. The researchers proposed a more direct methodology by incorporating Google Trends Index (GTI). They used keywords volume related to migration assuming people search before making a decision to move. A similar study was done by (Wladyka, 2017). The researcher evaluated the relationship between search terms reported by Google Trends (GT) and migration flows between Latin America and Spain. Both of these studies indicate the usefulness of GT as a predictor for migration flows. Nevertheless, this type of approach comes with drawbacks. Only a limited combination of keywords was included. In addition to that, the coverage of GT is restricted in underdeveloped countries, limited to one search engine type, only aggregated data for the volume of online search activity is available and just popular queries are included. These shortcomings can be diminished by making use of media reports that capture a variety of semantic links and are not restricted by the number of internet users in the country. Lin, Carley and Cheng (2016) extracted Goldstein ranking from the news articles provided by the GDELT. This information was used to construct the conflict and alliance network between country-pairs to build a model for migration movements between the countries. Ahmed et al. (2016) used the news articles provided by the GDELT in order to obtain the data about policy changes and other

occasions. This information was used to forecast a mass migration - refugee crisis in Europe by using various machine learning algorithms. The results have shown that the appropriate combination of selected methods and data can provide a potential relating to migration prediction. This assumption is further confirmed by Bijak et al. (2019) who provided the guidelines for migration forecasting. This involves three important steps: (1) understanding of attributes of migration fluctuations such as economic or political shocks; (2) assessing strengths, weaknesses, and uncertainty of available data on migration as there might be inconsistencies; (3) selecting the appropriate model for the type of migration.

Migration prediction is an important topic and there is an extensive literature devoted to analyzing the rationale behind the spatial mobility of people. Nevertheless, there is no clear set of predictors or a single methodology. Following the indications provided by (Bijak et al., 2019) and conclusions made from previous researches, in this thesis, it is assumed that various events concerning changes in migration can be detected in the news articles (Ahmed et al., 2016). Thus, media serves as a source to learn about the public sentiment which was found to have an effect on migration trends (Lumpe, 2018; Gorinas & Pytlikova, 2015). For these aforementioned reasons, the GDELT is chosen as it offers the largest near real-time information and provides a possibility to extract sentiment at the country-pair level. In terms of complete and reliable migration data, the focus is on New Zealand considering the most detailed, large scope, accurate data. Besides that, it provides a possibility to use distinct migration data rather than a net which is in accordance with (Raymer & Wisniowski, 2018). The machine learning methods have shown to perform better than the traditional models in numerous studies, especially when the data is complex (Robinson & Dilkina, 2018; Azose & Raftery, 2015). Therefore, a variety of different machine learning models is selected for comparison.

2. EXPERIMENTAL SETUP

2.1. DESCRIPTION OF DATA

As previously stated, the focus was on New Zealand considering the available data on migration was the most detailed, large scope and accurate. Thus, permanent and long-term migration data from the Statistics Department of New Zealand and the news sentiment extracted from the GDELT Event database were applied to build the new dataset for predicting migration flows in New Zealand. The connection between migration flow and the sentiment was examined at the country-pair level. That is, arrivals in New Zealand were studied in relation to sentiment towards 50 countries. In particular, this thesis analysed the arrivals for the countries that were reported in the top 50 list concerning immigration to New Zealand during the period of the last 18 years (New Zealand Immigration Statistics, n.d.). A full list of selected countries can be found in Table 1 in Appendix A: List of countries.

2.1.1. The Statistics Department of New Zealand database. The records for international travel and migration are gathered by the Customs Service from various airports and seaports in New Zealand. This information is collected regularly and provided to New Zealand's Statistics Department. Permanent and long-term migration is an actual count of totals by passenger type and direction (arrivals, departures) based on the details provided in the passenger card, flight and passport information (such as date of travel, date of birth and sex). The passenger type is based on their travel history and the time spent in and out of New Zealand. A traveler who arrives in New Zealand with an intention to stay for a minimum of 12 months is considered a long-term migrant (International Travel and Migration, Statistics New Zealand, 2018). Permanent and long-term migration by an every country of residence and citizenship dataset consists of time (monthly data, starting from April 1978 till December 2018), travel direction (arrivals, departures and net), country of residence (including countries and continents) and citizenship (distinguished between New Zealand, Australia, and the total)

(Table 1). Country of residence for arrivals is based on the last place of residence (where a person last lived for 12 months or more). The dataset is an open-source and is accessible from the Statistics New Zealand database in different formats (“.xls”, “.csv”, and “.tqx”) (Statistics New Zealand, n.d.).

Table 1.

Permanent & long-term migration in New Zealand by every country of residence and citizenship (monthly) dataset details.

Variables	Description	Selection
<i>Time</i>	Monthly data from April 1978 till December 2018	489 rows
<i>Travel Direction</i>	Arrivals, departures and net	3 columns
<i>Country of Residence</i>	All of the countries, continents and the total	3 columns
<i>Citizenship</i>	Separated between New Zealand, Australia and the total	3 columns
Total number of cells:		1 118 007

As mentioned previously the choice for this dataset was motivated by a high quality of migration data expressed in units. Available information on migration included a wide time period, separate arrivals data, and migration numbers for specific countries. For this thesis only three elements were selected: time (monthly data starting from 1978 04 till 2018 12), travel direction (Arrivals), country of residence (for a list of selected countries refer to Table 1 in Appendix A: List of countries).

2.1.2. The GDELT events database. The GDELT is an open-source database which monitors the world's news from around the globe and dates back to January 1, 1979, through the present day (Leetaru & Schrod, 2013). It is being continuously updated and includes more than 300 different categories of CAMEO-coded events. This database connects a huge amount of data about locations, organizations, persons, and events in over 100 different languages. It includes daily news from printed and online sources, as well as broadcasts from around the world (Elshendy & Fronzetti Colladon, 2017). The typical format of a raw event data includes but is not limited to the following elements: a time frame, fields indicating actors involved in the specific event (the first actor is a performer and the second - receiver), CAMEO code for event, an average tone (emotional connotation of the words in the article)

associated with that event (Delias, Zoumpoulidis, Kazanidis, 2019). An average tone element is used to identify the impact of the event by locating the positive and negative words in the article. It ranges from -100 (very negative) to +100 (very positive), although usually the score is between -10 and +10 (Leetaru & Schrod, 2013). At the aggregate country level, an average tone towards another country may reflect the public sentiment (Ye, 2017). The GDELT Event database records are accessible in “.csv” extension and are free to download (The GDELT Project, n.d.).

Table 2.
Selected variables from the GDELT Event database.

Variables	Description
“MonthYear”	Date for the event in “YYYYMM” format. Includes the time period from January 1, 1979 till December 31, 2018
“Actor1CountryCode”	The 3-character CAMEO code for the country associated with “Actor1” (“Actor1” corresponds to action initiator - New Zealand)
“Actor2CountryCode”	The 3-character CAMEO code for the country associated with “Actor2” (“Actor2” corresponds to action receiver – each of the 50 countries)
“AvgTone”	The average “tone” of all records including one or more mentions of this event during the 15-minute interval in which it was first detected. The average “tone” can be used as an estimate of the significance of an event and as a proxy for the “impact” of that event. It is based on the action performed by “Actor1” upon “Actor2”. The typical values range between -10 and +10.

Note. Descriptions for the variables from The GDELT Project (2015).

Potential reasons for decisions to emigrate from one country to another (such as policy changes, cooperation, bilateral agreements, to name a few) are captured under CAMEO-coded events. By selecting New Zealand as an action initiator (“Actor1”), the other countries as receivers (“Actor2”) and extracting an average tone between them for these events can provide a potential for better migration projection as some events are perceived as more negative or more positive. Thus, this corresponds to the sentiment between country-pair. The choice for this dataset was motivated by a variety of a real-time information, long time period, the possibility to aggregate data on a country-pair level, and to extract the average “tone” which was used to measure the sentiment between New Zealand and the selected 50

countries. For this thesis only four elements were selected (see Table 2): “MonthYear”, “Actor1CountryCode”, “Actor2CountryCode”, “AvgTone” (a full list of countries selected as “Actor2” can be found in Table 1 in Appendix A: List of countries).

2.1.3. Data pre-processing. The data pre-processing involves transforming the raw data into an organized format. This includes a few important steps: importing the datasets, checking for missing values, inspecting anomalies in the data and standardizing the variables.

2.1.3.1. Importing the datasets. The arrivals data containing the selected 50 countries was downloaded from the Statistics Department of New Zealand in “.csv” format for the period of April 1978 – December 2018 (Statistics New Zealand, n.d.). This dataset contained all the country names in each separate column and each month in a row. For the convenience country names and their number of arrival observations were restructured in rows. Afterward, country names were converted to country codes in order to combine this dataset with the GDELT dataset. After this initial processing, the arrivals dataset contained 24 450 observations. The GDELT was downloaded in “.csv” format datasets by filtering New Zealand’s country code (NZL) as “Actor1CountryCode” and the rest of the 50 countries as “Actor2CountryCode” for the time period of January 1979 – December 2018 (a list of countries with their country codes can be found in Table 1 in Appendix A: List of countries). This initial dataset contained 388 475 observations. Since arrivals data carries monthly observations, the variable representing sentiment was aggregated as a monthly average per country. After this, the number of observations was reduced to 10 009.

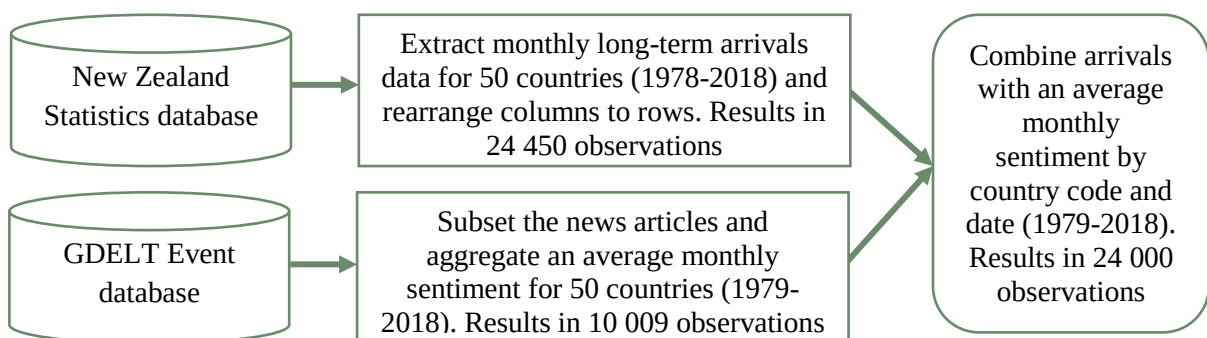


Figure 1. Overview of the databases and data preparation.

Afterward, both datasets were combined and filtered from January 1979 considering the first available date for the GDELT data. This combined dataset contained 24 000 observations and included 4 variables: date (year and month), country code, the number of arrivals and the average value for the monthly sentiment. An overview of the data preparation steps can be found in Figure 1.

2.1.3.2. Missing values. The arrivals variable did not contain any missing values for the selected 50 countries. However, the sentiment variable was missing a lot of observations (Figure 2). The majority of countries were missing 50 and even more percent of observations. Only Australia and the United States contained most of the observations for all of the period (1979-2018). Australia was missing 5 percent of data for monthly sentiment variable and the United States was missing fewer than 10 percent of data (Figure 2).

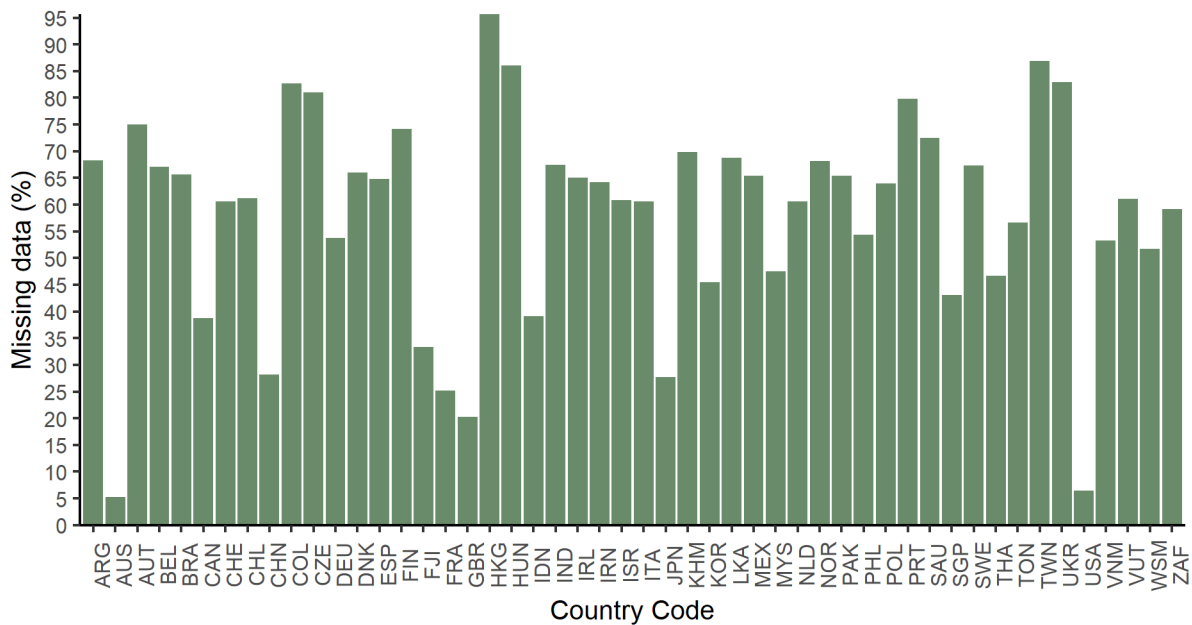


Figure 2. Overview of missing values for monthly tone variable per country.

Imputing that many sentiment values were not an option since the results would become more biased. Another plot was made to see how many months of data for sentiment are missing per year within each country (Figure 1 in Appendix B: Missing months per year). It revealed that most of data is contained in the recent years. In order to choose the biggest number of countries with the longest period available, it was decided to look into countries

having approximately 70 percent of data available for sentiment per year. That is, at least 8 months of sentiment data is available per year. Figure 3 displays all the countries and each black dot in the plot represents that at least 8 months of sentiment data is available per year.

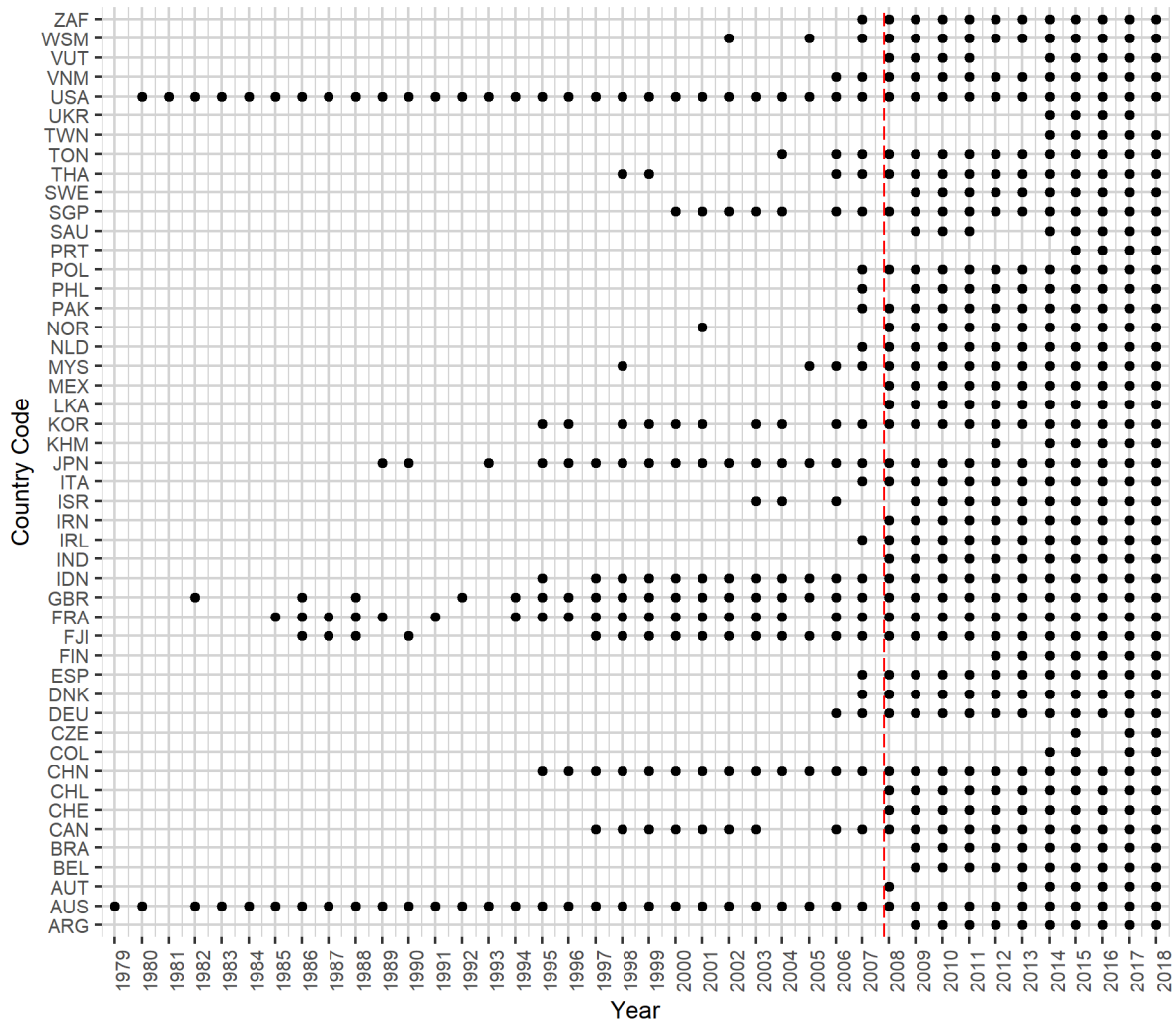


Figure 3. *At least 8 months of values for sentiment variable are available per year.*

Most of the data for sentiment variable was available from 2008 onwards. Therefore, 2008 was being chosen as a starting point (marked with the red line in Figure 3) for 32 countries (a full list of selected countries can be found in Table 1 in Appendix A: List of countries). However, to make use of the sentiment data available for almost all the years for two countries (Australia and the United States) it was decided to additionally select these countries starting from 1979 onwards. The starting point for Australia was the first monthly observation available regarding the sentiment variable which was April 1979. The starting

point for the United States was the second monthly observation available - which was July 1979 (due to the five consecutive missing observations between the first and the second data values). As a result, three subsets of data were formed: (1) the first containing the data with arrivals from Australia to New Zealand and the sentiment between these countries (New Zealand towards Australia) for the time period of April 1979 – December 2018; (2) the second containing data for the United States for the time period of July 1979 – December 2018 and (3) the third containing data for the 32 countries for the period of January 2008 – December 2018.

When data was not available, the approximation based on the average of the closest months was imputed by taking into account the previous and the next month's values. If there were consecutive months with missing values, the closest month to that missing value was selected. After missing data imputation the first subset resulted in 477 observations, the second subset resulted in 474 observations, the third subset resulted in 4 224 observations.

2.1.3.3. Anomalies in the data. A decrease in the monthly sentiment value was detected in the years 2012 and 2015 for all selected countries. To illustrate this decline, sentiment value for 4 countries (Canada, Great Britain, Japan and Fiji) is displayed in Figure 4. Although there is no official documentation regarding any changes in how an average tone is measured, the same issue was observed by other researchers. Kumar, Benigni and Carley (2016) highlighted that the change in “AvgTone” scale after March 2015 becomes mostly negative although before that positive values were seen. To correct for this decline, they reversed the observations before the date of the downturn. Weller and McCubbins (2014) also noted the absence of the negative values for an average “tone”. Following this, the sentiment value had to be normalized. In order to do that, z-scores were selected as a common normalization measure. They have a mean of 0 and a standard deviation of 1 (James, Witten, Hastie & Tibshirani, 2017). The mathematical formula is below:

$$Z_i = \frac{(x_i - \mu)}{\sigma}$$

Where x_i is observation, μ is the mean and σ is the standard deviation.

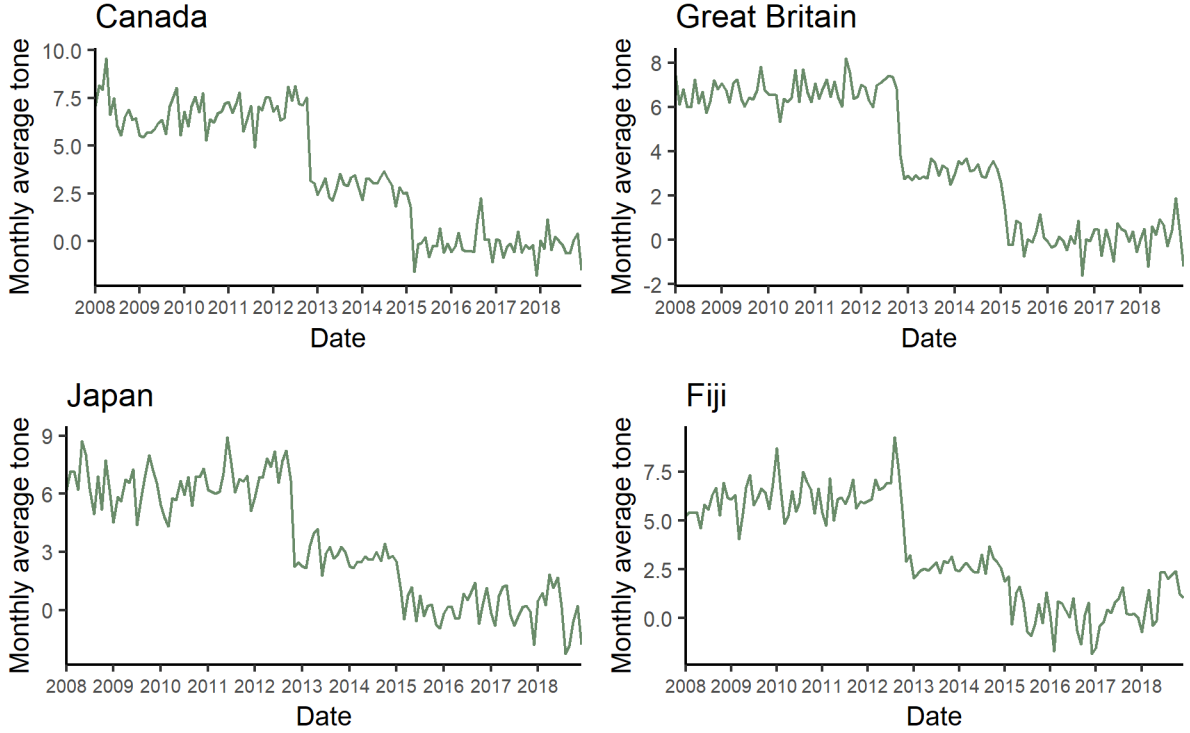


Figure 4. The monthly average tone visualization for 4 countries in 2008-2018.

The mean and the standard deviation were counted for all of the countries for each month per year and each value of the sentiment at the country-pair level was subtracted by the monthly mean for all the countries and divided by the monthly standard deviation for all the countries, which is defined as:

$$\frac{S_t^i - \bar{S}_t}{SD_t}$$

Where S_t^i is the sentiment observation for country-pair i at month t , $\bar{S}_t$ is the average value of the sentiment for all 32 country-pairs for the month t and SD_t is the sentiment standard deviation for all 32 country-pairs at month t .

After z-scores implementation, there was no more inconsistency in the sentiment data (Figure 5). For Australia and the United States subsets z-scores were performed in a slightly different manner due to the lack of monthly observations in the older years. For each of the

country, z-scores were calculated by averaging monthly tone observations per year and counting standard deviation per year. Meaning that $\bar{S}_t$ is the average value of the sentiment per year and SD_t is the yearly standard deviation. After the implementation of z-scores, the sentiment value ranged approximately between -2 and 3 (the sentiment value before and after normalization for Australia and the United States are displayed in Figure 1 and Figure 2, respectively in Appendix C: Sentiment before and after z-scores for Australia and the United States).

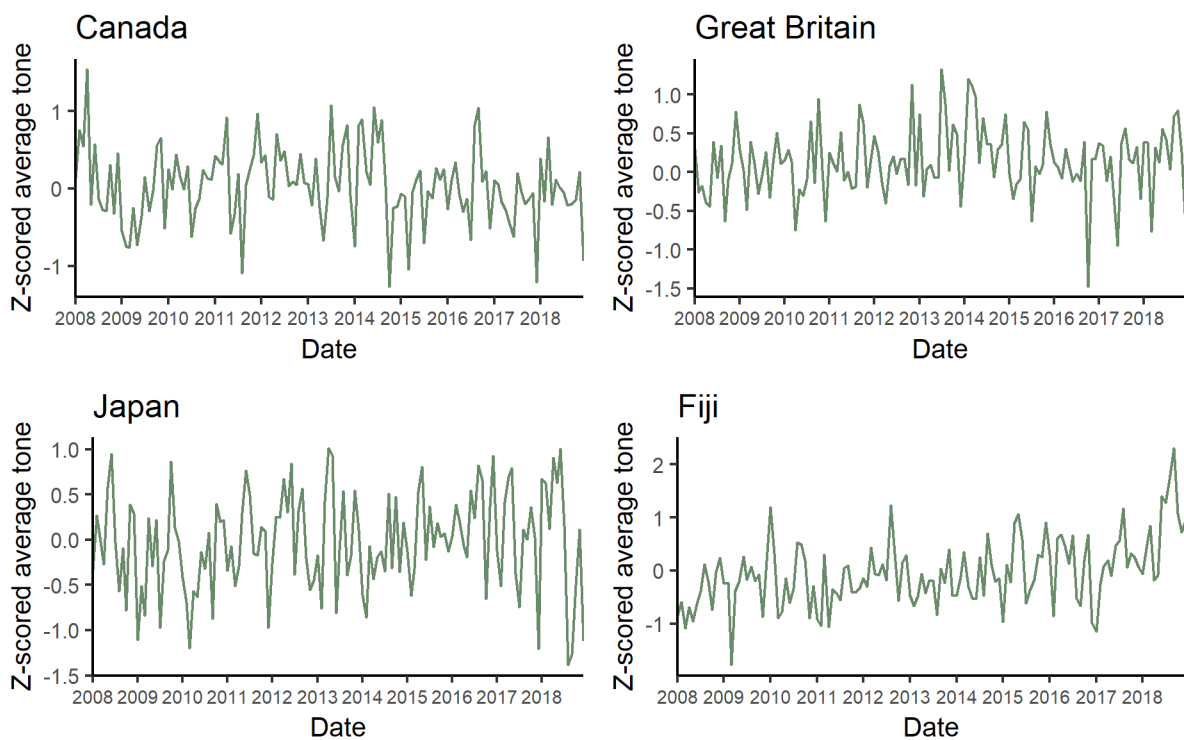


Figure 5. *The z-scored monthly sentiment illustrated for 4 countries in 2008-2018.*

In terms of arrivals data, the visual inspection of the time series in Figure 6 showed a strong seasonality pattern. After further investigation, the similar arrival patterns were found in all the countries being analyzed (Figure 1 in Appendix D: Seasonality detection). The biggest arrivals were in January and December, the smallest were in June. The number of arrivals averaged monthly for all of the countries in the time period of 2008-2018 showed a decreasing pattern till June and then increasing till December (Figure 2 in Appendix D: Seasonality detection). However a yearly seasonal plot displayed that it is not always the

case, for instance, a trend for 2018 do not follow the same pattern (Figure 3 in Appendix D: Seasonality detection). Therefore, month could be an important predictor and as a result, it was decided not to adjust seasonality but to use it as a feature.

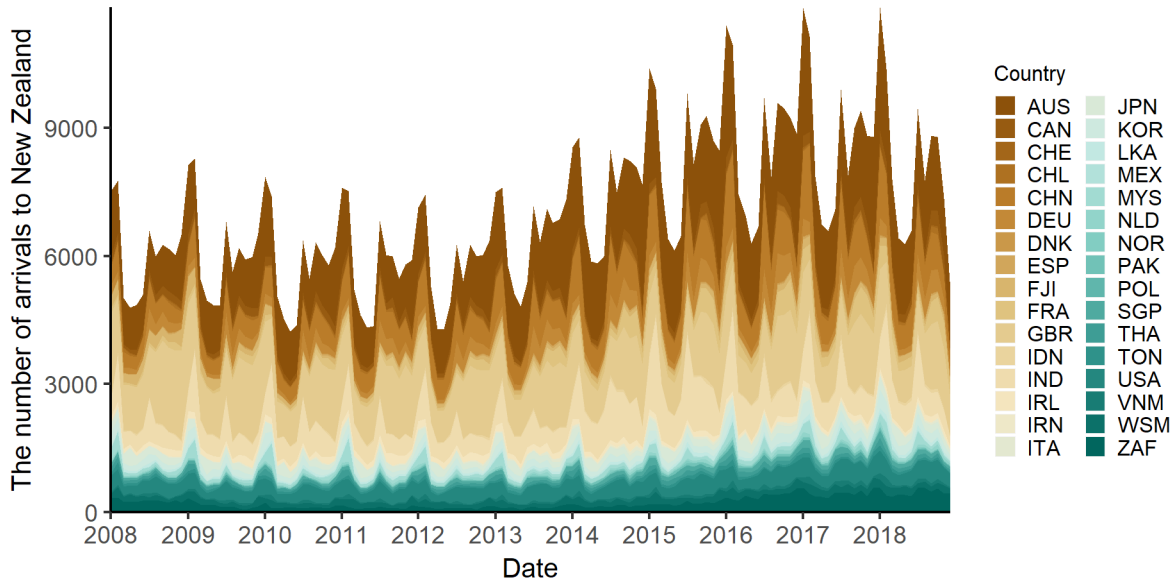


Figure 6. *The number of arrivals in New Zealand from 32 countries in 2008-2018.*

2.1.4. Sliding time window approach. Since arrivals and the sentiment are continuous time series variables, a supervised machine learning problem was formed or more precisely, regression prediction task. In order to apply machine learning algorithms to a time series data, a new dataset accounting for the sequential time series dependency had to be constructed. The straightforward process of dealing with the long term forecasting is to divide it into a set of smaller independent prediction tasks (Bontempi & Taieb, 2011). This can be achieved by the sliding time window approach. The procedure behind the sliding time window approach is as follows: using a full sample period with the fixed number of observations of consecutive time series, construct new subsamples by shifting forward one observation at a time (Galloppo, 2010). The number of observations used to create each subsample is referred to as the size of a window which is determined by a researcher (Inoue, Jiny & Rossi, 2017). A large window is not able to capture parameters movements in the subsamples and a small one

provides detailed changes but variation is much larger (Nakajima, Matsui, Sakai & Yagi, 2011).

Figure 7 illustrates the sliding time window approach. A Window 1 contains information on input and output observations for the first three months. A Window 2 is created by moving one-month forward. The information captured in this window contains the last two observations from Window 1 and new observations for input and output for the next month (month four). Each consecutive window is slid one-month ahead until there are no more observations in the time series dataset. This way a sequential time series pattern is accounted for as observations in each subsample overlap.

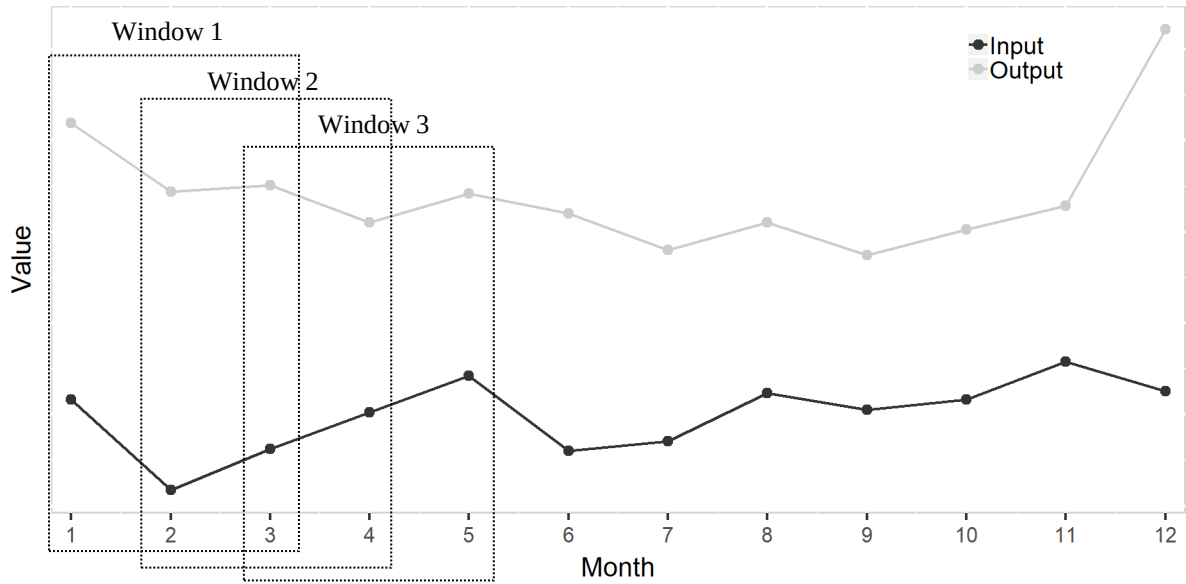


Figure 7. Illustration of the sliding time window approach.

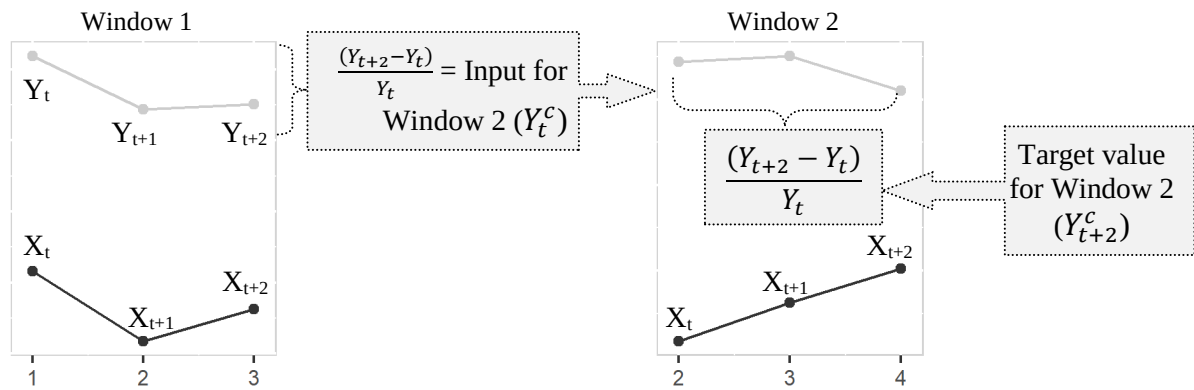


Figure 8. A close-up for Window 1 and Window 2.

For instance, consider the sliding time window of a size three. Independent variable X and dependent Y are showed for different time periods ($t, t+1, t+2$) in a close up of Window 1 (Figure 8), where t represents the first month in a time series data. In order to predict Y_{t+2} value, the model is fed with the information available on X_t, X_{t+1} and X_{t+2} , also the dependent variable Y_t . For the purpose of this thesis instead of the absolute arrivals' numbers, a relative monthly change in arrivals two months ahead was assessed. In other words, a dependent (target) variable in Window 2 was defined as: $\frac{(Y_{t+2}-Y_t)}{Y_t}$ (or Y_{t+2}^c , where c is used to denote the change) and the former relative monthly change in arrivals from Window 1 (Y_t^c) was used as a starting input (Figure 8). Independent variable (the sentiment) was included for all three months (X_t, X_{t+1}, X_{t+2}) in each time window. The transformed time series data was used in the following analysis.

In order to evaluate the models' performance in each prediction task, a simple method that serves as a baseline had to be established. For this reason, it was assumed that the monthly relative change in arrivals for a previous month (Y_t^c) was the same as the monthly relative change in arrivals two months ahead (Y_{t+2}^c). That is, the target value for Window 2 was expected to be equal to the input value (output value from Window 1).

2.1.5. Additional features. In order to increase the amount of data available for the prediction task in each time window, additional features were extracted from arrivals and sentiment data. In addition to that due to the seasonality detection in arrivals data monthly interactions regarding a relative monthly change in arrivals were constructed as features too. Table 3 displays a summary of the extracted and additional features in the sliding time window dataset.

Table 3.
Extracted features and their description.

<i>Feature</i>	<i>Description</i>	<i>Type</i>
A_t	Absolute arrivals value at month t	Numerical
A_{t+1}	Absolute arrivals value at month $t+1$	Numerical
A_{t+2}	Absolute arrivals value at month $t+2$	Numerical

Table 3 (continue).

A_{rc}	Relative monthly change in arrivals $\frac{(A_{t+2}-A_t)}{A_t}$	Numerical
S_t	Absolute sentiment value at month t	Numerical
S_{t+1}	Absolute sentiment value at month $t+1$	Numerical
S_{t+2}	Absolute sentiment value at month $t+2$	Numerical
S_{mean}	Average sentiment value $\frac{(S_t+S_{t+1}+S_{t+2})}{3}$	Numerical
S_{max}	Maximum sentiment value of the set of S_t, S_{t+1}, S_{t+2}	Numerical
S_{min}	Minimum sentiment value of the set of S_t, S_{t+1}, S_{t+2}	Numerical
S_{median}	Median value in the sorted list of the set of S_t, S_{t+1}, S_{t+2}	Numerical
S_{sd}	Spread between sentiment $\sqrt{\frac{\sum (S_{mean}-S_n)^2}{n-1}}$	Numerical
S_{var}	Variance of sentiment $\frac{\sum (S_{mean}-S_n)^2}{n-1}$	Numerical
S_{slope}	Direction and the steepness of the sentiment line $\frac{S_{t+2}-S_t}{2}$	Numerical
S_{dir}	Direction of the slope: 1 if slope is going up, -1 - going down, 0 - is stable	Binary
S_{range}	Difference between the maximum and minimum sentiment value $S_{max} - S_{min}$	Numerical
<i>MarJan</i>	Monthly interaction between March and January. 1 if present, 0 otherwise	Binary
<i>AprFeb</i>	Monthly interaction between April and February. 1 if present, 0 otherwise	Binary
<i>MayMar</i>	Monthly interaction between May and March. 1 if present, 0 otherwise	Binary
<i>JunApr</i>	Monthly interaction between June and April. 1 if present, 0 otherwise	Binary
<i>JulMay</i>	Monthly interaction between July and May. 1 if present, 0 otherwise	Binary
<i>AugJun</i>	Monthly interaction between August and June. 1 if present, 0 otherwise	Binary
<i>SepJul</i>	Monthly interaction between September and July. 1 if present, 0 otherwise	Binary
<i>OctAug</i>	Monthly interaction between October and August. 1 if present, 0 otherwise	Binary
<i>NovSep</i>	Monthly interaction between November and September. 1 if present, 0 otherwise	Binary
<i>DecOct</i>	Monthly interaction between December and October. 1 if present, 0 otherwise	Binary
<i>JanNov</i>	Monthly interaction between January and November. 1 if present, 0 otherwise	Binary
<i>FebDec</i>	Monthly interaction between February and December. 1 if present, 0 otherwise	Binary

2.2. METHODOLOGY

2.2.1. Description of experiments. The main purpose of this thesis was to predict migration flow (arrivals) relative monthly changes in New Zealand from other countries based on the sentiment value between New Zealand and other countries. The interest was in predicting two months ahead relative monthly change in arrivals (Y_{t+2}^c) based on the previous change (Y_t^c). For all of the experiments, a new subset of data was created by the sliding time window approach with a window width of three. After this implementation the goal in each window was to predict the future value (Y_{t+2}^c) by using sentiment values for a three-month

period (X_t, X_{t+1}, X_{t+2}) and additional features, also the previous value of monthly relative change (Y_t^c) was included.

2.2.1.1. Experiment 1. How accurately do commonly used models predict migration flow to New Zealand and how can they be improved? Do predictive models generalize?

The goal for the research question 1 was to compare various machine learning algorithms and their predictive power by using the first subset of data containing arrivals to New Zealand and sentiment from one country – Australia for a time period of April 1979 till December 2018. In addition to that, generalization ability of predictive models and their tuning parameters was tested on another country – the United States for the time period of July 1979 till December 2018 (data contained in the second subset).

2.2.1.2. Experiment 2. Are there any migration patterns within different continents in terms of migration flow to New Zealand? Is there a single best performing algorithm?

The goal for the research question 2 was to predict relative monthly changes in arrivals to New Zealand from the countries in continents by using the same algorithms as in research question 1. This way patterns, if there were any, were found within continents. Also, it provided an opportunity to find the best performing model in different scenarios. In order to conduct this experiment, a window procedure was performed for each country from a total of 32 for a time period of January 2008 – December 2018 (contained in the third subset). Afterward, country windows were grouped into 5 continents: Oceania, Europe, Asia, Americas, Africa and the Middle East (see Table 4). Since Africa and the Middle East continent contained only two countries, it was decided to exclude it from further research. For the rest of the continents, one country per continent was selected as a test. Tonga, Denmark, India, and Mexico (Table 4 in bold text) for Oceania, Europe, Asia and Americas respectively each containing 129 observations. The training set for Oceania and Americas contained 387

observations for three countries; the training set for the rest of the two continents contained 1290 observations for ten countries.

Table 4.
Countries grouped into continents.

<i>Continent</i>	<i>Country name</i>	<i>Number of observations</i>
<i>Oceania</i>	Australia, Fiji, Western Samoa, Tonga	Train: 387 & Test: 129
<i>Europe</i>	Netherlands, Spain, Italy, Switzerland, Ireland, Germany, Great Britain, Poland, Norway, France, Denmark	Train: 1290 & Test: 129
<i>Asia</i>	Vietnam, Indonesia, Thailand, Singapore, Malaysia, Korea, China, Sri Lanka, Pakistan, Japan, India	Train: 1290 & Test: 129
<i>Americas</i>	The United States, Canada, Chile, Mexico	Train: 387 & Test: 129
<i>Africa and the Middle East</i>	South Africa, Iran	Total: 258

Note. Countries in bold text were selected for a test set and were excluded from a train set.

2.2.1.3. Experiment 3. Is the sentiment an important determinant in migration prediction?

The aim of the research question 3 was to evaluate the importance of the sentiment as a determinant of arrivals forecast in predictive models. This was based on extracting variable importance from the best performing models in research questions 1 and 2.

2.2.2. Models and tuning parameters. In order to choose the most suitable model for the prediction task and for the sake of comparison, various machine learning models were selected. Most of the algorithms, for instance, Lasso, Ridge, Linear Regression, Gradient Boosting of Regression Trees, Support Vector Machine were used in numerous researches concerning migration prediction (Ahmed et al., 2016; Robinson & Dilkina, 2018; Bohme, Groger & Stohr, 2018). Each model parameter was tuned via 10-fold cross-validation to achieve the best performance (models with the main tuning parameters can be seen in Table 5).

A standard linear regression is used to describe the relationship between a set of predictors and a response variable. The model is fit using least squares (James, Witten, Hastie, Tibshirani, 2017). It is a very simple method for numerical predictions and serves as an elementary unit for more complex models. It is formally defined as:

$$x = w_0 + w_1 a_1 + w_2 a_2 + \dots + w_k a_k$$

Where x is the response variable, $a_1, a_2, \dots, a_k$ are the attribute values and $w_0, w_1, \dots, w_k$ are weights (Witten & Frank, 2009).

Ridge introduces regularization by giving some variables smaller weights and pushing them towards a zero value, which is achieved by an L2 penalty. An additional parameter, lambda, is introduced which controls the model's complexity. When lambda is large, the model is complex and all the variables are pushed towards zero. When lambda is close to 0, the model performs ordinary least squares (OLS) (Conway & White, 2012). Ridge has an advantage over OLS to prevent overfitting by reducing the variance (with a small increase in bias) of the predictors when the lambda increases. It is formally defined as:

$$\sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p \beta_j^2 = RSS + \lambda \sum_{j=1}^p \beta_j^2$$

Where λ corresponds to lambda, $\lambda \sum_{j=1}^p \beta_j^2$ is a shrinkage penalty, $\beta_1, \dots, \beta_p$ are coefficient estimates (James, Witten, Hastie & Tibshirani, 2017).

Lasso is an alternative method for Ridge. Instead of an L2 penalty in Ridge, Lasso uses an L1, which has a possibility to force some of the variables to be exactly zero when the tuning parameter – lambda, value is sufficiently large. In that way, Lasso can be used as a variable selection method (James, Witten, Hastie & Tibshirani, 2017).

When building a Random Forest a random subset of predictors is chosen from the full set each time when a split in a tree is considered. Usually, the number of predictors at each split is equal to the square root of the total number of predictors. The number of trees being build on bootstrapped training samples has to be sufficiently large for the error rate to settle down (James, Witten, Hastie & Tibshirani, 2017). Random Forests rely on four parameters: the subsample size, the number of trees, the number of selected features in each split and the tree depth. The best performance is obtained with a large number of trees in the forest but the

optimal number is achieved when the minimum error cannot be reduced even with an increasing number of trees. The tree depth and the subsample size have a similar impact to a Random Forest performance, so it is sufficient to tune either of these (Scornet, 2018).

Support Vector Machine (SVM) is well suited for time series prediction. When using it for a regression task, cost function C must be introduced as a way to measure the empirical risk to minimize the error (Boyle, 2011). C represents the trade-off between the complexity of the model and the approximation error, ϵ (Lahiri & Ghanta, 2008). ϵ is another parameter which represents the loss function. It affects the number of support vectors selected. The bigger value of this parameter suggests a fewer number of support vectors selected and more consistent predictions (Cherkassky & Ma, 2004). The performance of SVM is highly affected by a good set of the aforementioned parameters as well as kernel selection (Lahiri & Ghanta, 2008).

The main idea of Gradient Boosting Machine (GBM) is fitting an additive model to a group of models sequentially. At each step, a new model is trained with respect to an error of a whole group of models learned so far. An advantage of using boosting is that it complies with both - variance and bias (Natekin & Knoll, 2013). Some important parameters in this model are the number of trees, tree depth, learning rate and the number of samples. The idea is to discover the optimal number of trees because the model can easily overfit. This risk can be reduced by using a smaller learning rate and only a proportion of the training sample instead of a full size. The complexity of a model is controlled via tree depth (Boehmke, 2018).

All of the models were built and tuned in R Studio (version 3.5.1). A full list of packages used can be found in Appendix E: Libraries.

Table 5.
Models and the main tuning parameters.

<i>Models</i>	<i>Parameter</i>
Linear Regression	none
Ridge	λ (lambda), α (alpha)
Lasso	λ (lambda), α (alpha)
Random Forest	Subsample size Number of trees Number of features in each split Tree depth
Support Vector Machine	C (Cost) ε (loss function) Kernel parameters
Gradient Boosting Machine	Learning rate (shrinkage) Number of trees Tree depth

2.2.3. Evaluation metrics. Any prediction method had to be compared to the baseline to ensure that the new models perform better than this simple method. As previously mentioned, a Naive model was chosen which sets all forecasts for any period to be equal to the previous period actual observation (Hyndman & Athanasopoulos, 2018). That is, the value of a relative monthly change in arrivals two months ahead (Y_{t+2}^c) was considered to be the same as the value of the previous change (Y_t^c), which was defined as:

$$Y_{t+2}^c = Y_t^c$$

To evaluate the performance of different methods and to choose the best one, the data was randomly split into the training and testing sets by 80/20 ratio, respectively. Moreover, the data in the training set and test set was shuffled so it contained a random order of countries and different time frames in each window. The true prediction capability was measured on an independent test set containing unseen observations. A training error serves as a tool for comparing different methods but it is not a good estimate of the test error as it decreases when a model complexity increases. In an extreme case – when the training error is

equal to zero – model is overfitting which leads to a very high test error (Hastie, Tibshirani & Friedman, 2009).

Once the data was split into the train and test sets, the performance of a model was measured by an error term. The Mean Squared Error (MSE) is one of the most common error evaluation techniques (Witten & Frank, 2009). An optimal solution is found when MSE value is minimized, that is when predictions are close to the true value. MSE is formally defined as:

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{f}(x_i))^2$$

Where $\hat{f}(x_i)$ is the prediction that $\hat{f}$ gives for the i th observation (Hastie, Tibshirani & Friedman, 2009).

Cross-validation serves as a method to prevent overfitting. It is used to estimate the tuning parameters and to evaluate a forecast error of the ultimate model (Conway & White, 2012; Hastie, Tibshirani & Friedman, 2009). The most common way is to use 10-fold cross-validation because numerous tests showed that it is the right number to get the best estimate of an error (Witten & Frank, 2009). The procedure is as follows: training data are randomly divided into 10 parts; each part is kept in turn for validation and the models are trained on the remaining 9 parts (Table 6). The error rate is calculated on the kept part. This procedure is repeated 10 times and the error is averaged over 10 error estimates from the validation parts (Witten & Frank, 2009).

Table 6.
Algorithms and their tuning parameters.

	<i>Fold</i>									
	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>	<i>9</i>	<i>10</i>
Split 1	Validate	Train	Train	Train	Train	Train	Train	Train	Train	Train
Split 2	Train	Validate	Train	Train	Train	Train	Train	Train	Train	Train
....										
Split 10	Train	Train	Train	Train	Train	Train	Train	Train	Train	Validate

3. RESULTS

3.1. RESEARCH QUESTION 1

The aim for the research question 1 was to predict a monthly relative change in arrivals two months ahead (Y_{t+2}^c) from Australia to New Zealand based on the sentiment between these two countries. The sliding time window of the window width of three months was used to build a new dataset and to extract the temporal features (features are listed in section 2.1.5.). Afterward, the data was randomly separated into a training set containing 80 percent of the data (379 observations) and the remaining 20 percent was set as a test set (95 observations). Six different regression models were used in this experiment and their hyperparameters were tuned using 10-fold cross-validation (regression models and their tuning parameters can be found in section 2.2.2). Each model was evaluated against the training and test sets by calculating MSE value and for the sake of comparison, the Naïve baseline model was used. All the models, their best parameter values, training, and test MSEs are listed in Table 7 (best performing model is noted in bold text). Some more detailed information on models' tuning process can be found in Appendix F: Results of experiment 1.

Linear Regression was used first. Since linear models are highly affected by a risk of multicollinearity, features that are highly correlated were removed (S_{mean} , S_{median} , S_{slope} and monthly interaction FebDec) before fitting the model. Ridge model was tuned with 100 different values of lambda and the default value equal to 0 was selected for alpha. The model performed the best with the lambda value of 0.0235 which was very close to a 0 value, meaning that the model was very similar to a Linear Regression. Lasso parameter lambda was also tuned with 100 different values and the default value equal to 1 was selected for alpha. The best lambda value minimizing the MSE was 0.0013, which was even closer to a 0 value compared with Ridge. The Random Forest model was firstly tuned with 500 trees. 482 trees were returned as an optimal solution which minimized MSE. In order to reduce MSE

value, even more, the model was fed with 100 different combinations of the sample size and the number of features to be selected for training the model. The features were varied from 1 to 25 and the sample size in the interval of (0.55, 0.63, 0.80, 1.00). The best number of features was found to be 15 based on full sample size.

Table 7.
Models performance and their best tuning parameters.

<i>Models</i>	<i>Hyperparameters</i>	<i>MSE train</i>	<i>MSE test</i>
Baseline	-	0.1301	0.0945
Linear Regression	-	0.0172	0.0165
Ridge	$\lambda = 0.0235$, $\alpha = 0$ (default)	0.0198	0.0271
Lasso	$\lambda = 0.0013$, $\alpha = 1$ (default)	0.0197	0.0156
Random Forest	Subsample size = 1 Number of trees = 482 Number of features in each split = 15 Tree depth = 5 (default)	0.0181	0.0189
SVR	$C = 1.4142$ $\varepsilon = 0$ Kernel = Radial Learning rate = 0.1	0.0246	0.0188
GBM	Number of trees = 93 Tree depth = 3 Number of observations in nodes = 5 Fraction of the training set = 1	0.0107	0.0186
		Average MSE	
		0.0343	0.0300

Support Vector Regression was firstly fit with the cost parameter of 1, epsilon 0.1 and 165 number of support vectors. This means that approximately 45 percent of the training data was used. To increase the performance of the model it was tuned with a radial kernel and varying epsilon and cost hyperparameter values. The epsilon value was tuned in a range of 0 to 1 each time increasing its value by 0.1. The cost hyperparameter was tuned in a range of $2^{0.5}$ and 2^9 . The best results were achieved with an epsilon value of 0 and the cost value of 1.4142. Gradient Boosting Machine was firstly trained with 10 000 iterations (trees), the complexity of the model (interaction depth) of 1 and the learning rate (shrinking value) of 0.001. The optimal performance was achieved with 9 999 iterations. Afterwards the

hyperparameters were tuned by varying the learning rate in a range of 0.01, 0.1 and 0.3; the complexity in a range of 1, 3 and 5; the minimum number of samples to be used in a node to perform a split in a range of 5, 10 and 15; the sample size to be used in a range of 0.65, 0.8 and 1. This created 81 different combinations of hyperparameters to tune. The best result was achieved with a learning rate of 0.1, the complexity of 3, 5 samples to be used in a node, trained on a full data set with 93 iterations.

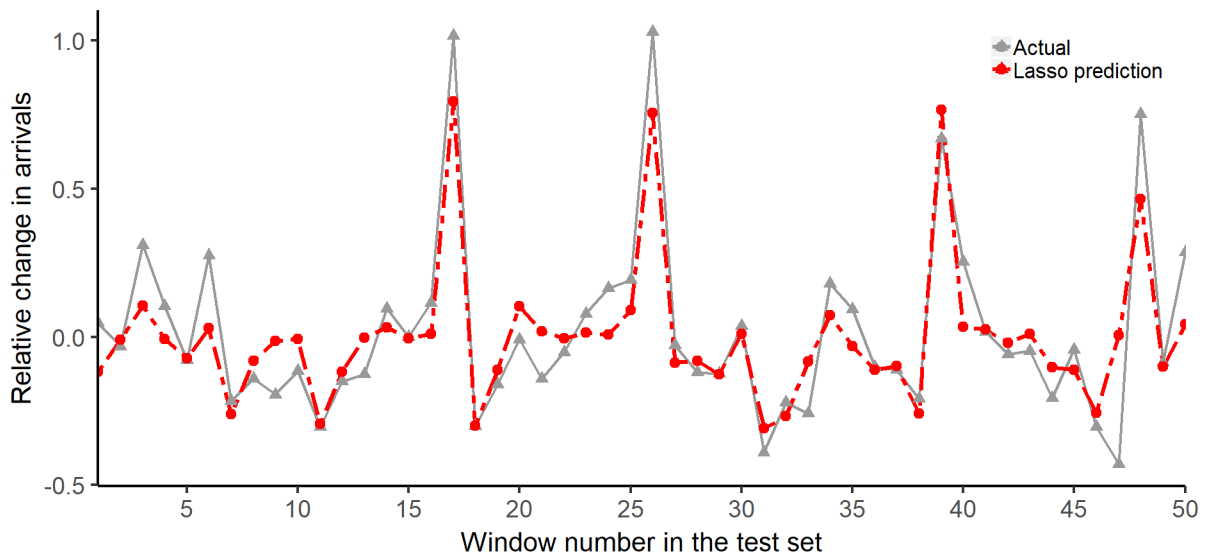


Figure 9. *Lasso predictions and the actual values on a test set for the first 50 observations.*

From Table 7, it is observed that all the models were able to capture the relative monthly changes in arrivals fairly well compared with the baseline model. The average MSE on a test set for all the models was 0.300 while the baseline MSE was 0.0945. The tree-based models and support vector machine produced almost identical MSE values on a test set. On average, these models were off by approximately 0.0188 from real value. However, the best performing model was Lasso. The MSE value for this model was 0.0156 in the test set, meaning that on average the model is off by this number from an actual relative monthly change in arrivals. This value suggested quite an accurate result in comparison to other models and the baseline. Lasso was performing approximately 16.5 percent better than the baseline. Figure 9 displays real values of the monthly relative changes in arrivals and Lasso predictions on a test set for the first 50 observations. The result of this visualization was

encouraging – Lasso predictions were quite close to the real values. This was in accordance with a small value of MSE in the test set. The worst performing model – Ridge MSE was 0.0271 which is approximately 1.15 percent worse than the best performing model.

The second part of the research question 1 was to assess the predictive power of the models' ability to generalize on unseen data. Regression models with their tuning parameters trained on arrivals and sentiment data from Australia were used to predict the changes in migration flow from the United States to New Zealand.

Table 8.
Models generalization performance.

<i>Models</i>	<i>MSE train</i>	<i>MSE test</i>
Baseline	0.1301	0.2360
Linear Regression	0.0172	0.1738
Ridge	0.0198	0.1508
Lasso	0.0197	0.1698
Random Forest	0.0181	0.1612
SVR	0.0246	0.1609
GBM	0.0107	0.1627
Average MSE		
	0.0343	0.1736

Note. The MSE train values are from training the models on Australia.

All the models were able to generalize to some extent since they outperform the baseline model (Table 8). Despite that, the increase in the test MSE for the United States data in contrast to the test MSE for Australia increased by ten units. This increase was remarkably high by taking into consideration how small in general the value of MSE was. The average test MSE for all the models trained on Australia was 0.0300 while the average test MSE for the United States was 0.1736. This indicates that the models could not perform equally well when the data came from a different distribution (another country). However, the baseline test MSE value for the United States compared with the baseline test MSE for Australia had also increased. The difference between these two MSEs was 0.1415 or almost 250 percent more. Analogously, a very similar increase in test MSE was noted for the rest of the models (the average difference was 0.1440). This suggested that there might be more extreme values

involved in the United States arrivals to New Zealand compared with arrivals from Australia and it could be one of the reasons why models were performing worse.

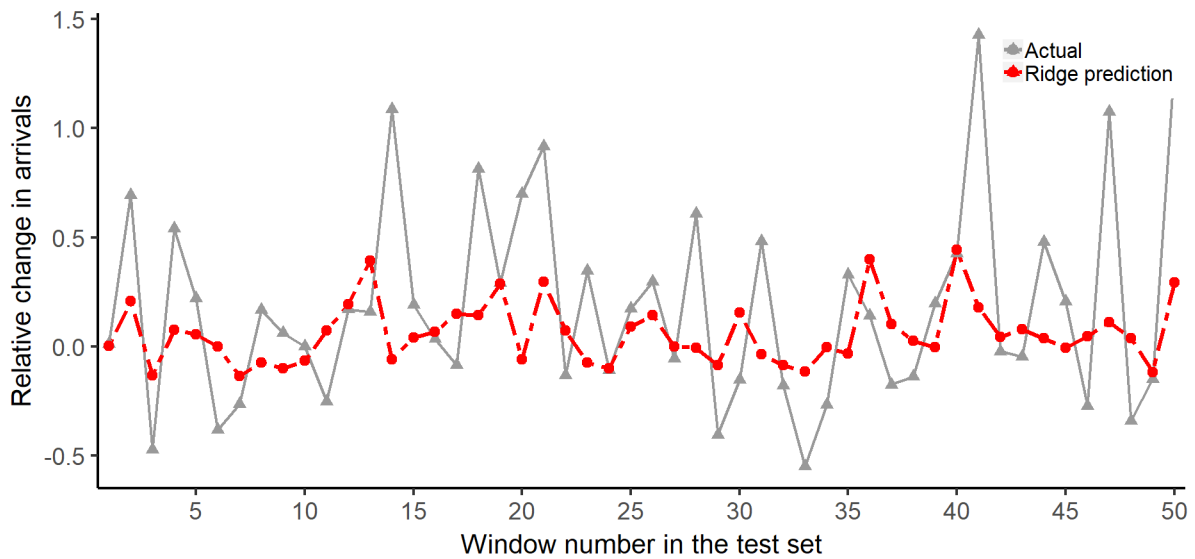


Figure 10. Ridge predictions and the actual values on a test set for the first 50 observations.

Worth mentioning is that Ridge was the worst performing model in terms of forecasting relative monthly changes in arrivals from Australia two months ahead in the first part of experiment 1. Surprisingly the test MSE for this model was the smallest among all the models in the second part of the experiment (fitting the models on arrivals and sentiment data from the United States). The test MSE for the baseline model was 0.2360 while Ridge was performing almost 64 percent better. Figure 10 shows Ridge predictions and the actual values only for the first 50 observations in the test set. As can be seen in this figure, the forecasts provided by Ridge for relative monthly changes in arrivals from the United States to New Zealand were considerably more uncertain than the ones provided by Lasso for the changes involving arrivals from Australia. In addition, it confirms the assumption that arrivals from the United States varied more than arrivals from Australia (the scale for the changes in arrivals was approximately between -0.7 to 1.5 compared with -0.4 and 1 , respectively). Despite that, most of the models' MSEs were smaller than the average MSE for all the models (0.1736) as well as the baseline (0.2360). This indicated that models can generalize to some extent although not all of the parameters might be suitable for this new data.

3.2. RESEARCH QUESTION 2

The aim of the research question 2 was to predict a relative monthly change within countries in the same continents, discover any patterns and generalize on a test set (the held-out country). The models were trained on the data containing Oceania, Europe, Asia, and Americas countries and tested on a held-out country (a list of countries involved in this experiment can be found in Table 3 in section 2.2.1.). The models' tuning procedure was similar to the one described in the research question 1 and only results are discussed for this research question. The MSE values for training and testing sets are displayed in Table 9.

Table 9.
Algorithms performance in continents.

<i>Method</i>	<u>OCEANIA</u>		<u>EUROPE</u>		<u>ASIA</u>		<u>AMERICAS</u>	
	<i>MSE train</i>	<i>MSE test</i>	<i>MSE train</i>	<i>MSE test</i>	<i>MSE train</i>	<i>MSE test</i>	<i>MSE train</i>	<i>MSE test</i>
Baseline	0.2763	0.8212	10.8190	11.9201	1.5427	2.0554	1.3258	4.9822
Linear Regression	0.0864	0.3557	5.1481	4.9422	0.4715	0.4681	0.6421	2.9566
Ridge	0.0946	0.3934	5.3250	6.0746	0.4950	0.6157	0.7054	2.7986
Lasso	0.0943	0.3460	5.3228	5.0696	0.4938	0.4762	0.7100	2.8333
Random Forest	0.0994	0.3359	5.2478	5.7550	0.5108	0.6112	0.7107	2.7832
SVR	0.1107	0.3957	5.6031	6.5764	0.5368	0.7013	0.7091	2.8043
GBM	0.0509	0.3730	3.2124	5.6296	0.4337	0.5141	0.4862	2.9188
Average MSE								
	0.1161	0.4316	5.8112	6.5668	0.6406	0.7774	0.7556	3.1539

Starting with the initial performance of all the models, it was seen that all of them outperform the baseline model. However, there was no single model that performs the best in terms of all the continents. Worth mentioning is that the Random Forest was the best performing model in Oceania and Americas which had the smaller training size (387 observations) and Linear Regression reached the smallest MSE in Europe and Asia where a training set size was much larger (1290 observations). This suggested that with more countries included in the training set models cannot make a difference between country

specifics and common patterns within a continent. That could be one of the reasons why Linear Regression was outperforming other models in the aforementioned continents. Also, a huge baseline MSE in Europe (10.8190 in the training and 11.9201 in the testing set) compared with the baseline MSEs in other continents (ranging from 0.2763 to 1.5427 and from 0.8212 to 4.9822 in the training and testing sets respectively) suggested that countries within this region were very different in terms of arrivals in New Zealand. Analogously, this increase in MSE value was seen in all of the models within Europe.

Since Americas continent included North, Central, and South American countries the models also did not perform well in the test set compared with the training set. An average MSE for all the models was 0.7556 in the train set and 3.1539 in the test set. The difference between the train and test MSEs in Americas was the greatest. The test MSE was more than 4.1 times bigger than the train MSE in contrast to 3.7, 1.1 and 1.2 for Oceania, Europe and Asia, respectively. This also suggested how much arrival patterns differ within the countries in different continents. Perhaps smaller regions could provide better predictions.

The most similar countries in terms of arrivals were within Asia region. Even with ten different countries included in the train set the average train MSE was 0.6406 in contrast to Europe region that included the same number of countries (train MSE was 5.8112). Also, the increase in the test MSE was the smallest among the continents. The difference between the test and train error was 0.1368. This could be an indicator of a general trend of arrivals from Asia to New Zealand.

Figure 11 demonstrates predictions made by the best performing models in the aforementioned continents only for the first 30 observations in a test set. Americas and especially Europe continent had a considerably bigger scale of relative monthly changes in arrivals compared with the rest of the continents, as expected. A relative monthly change in arrivals varied approximately between -0.5 and 10 for these continents while in Oceania and

Asia the amplitude was between -1 and 4 . This could be one of the reasons why models produced bigger MSEs within particular continents.

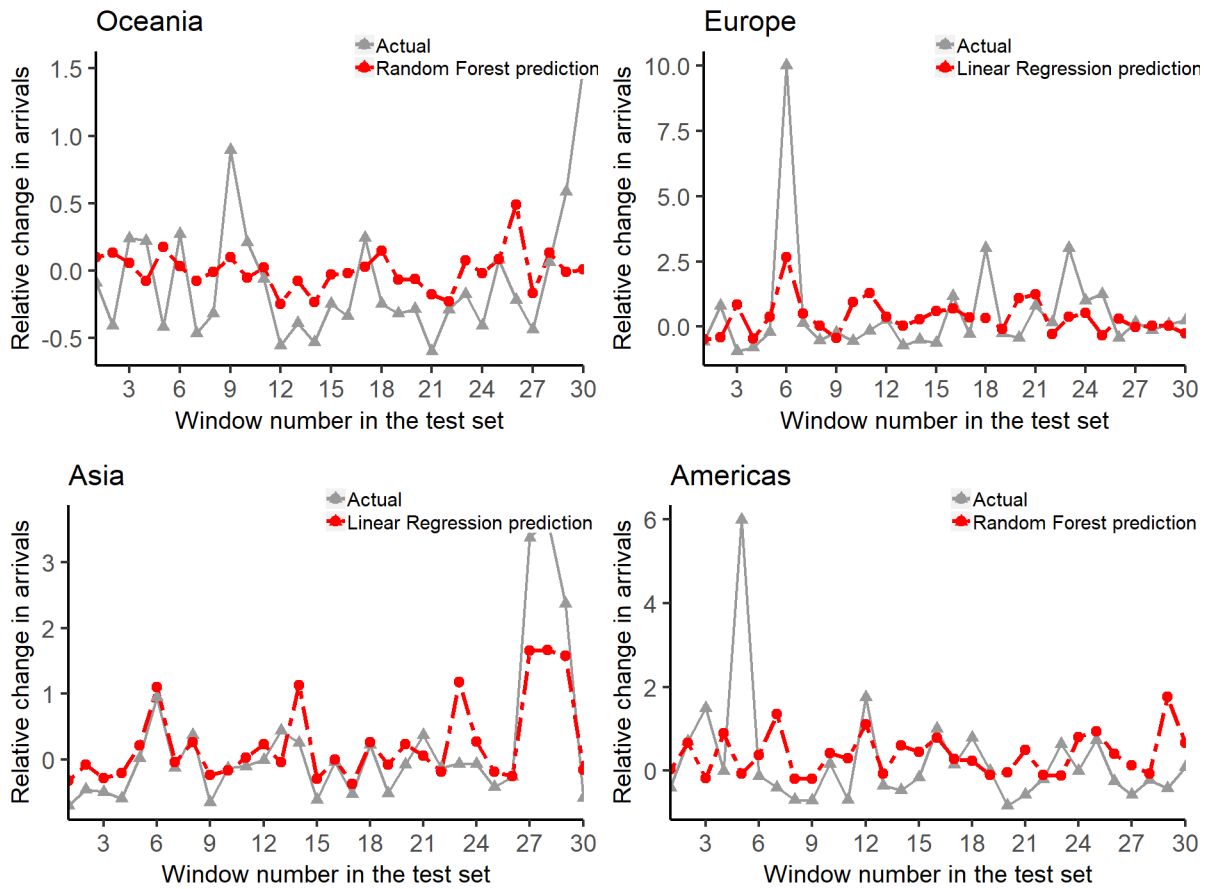


Figure 11. *The models' predictions and actual values for the first 30 observations in a test set.*

3.2. RESEARCH QUESTION 3

The goal of the research question 3 was to evaluate the importance of sentiment in predicting a relative monthly change in arrivals. To examine the possible impact the best performing models in research question 1 and 2 together with their feature importance were inspected (the exact coefficient values for variable importance can be found in Table 1 in Appendix G: Results of experiment 3).

Starting from the first experiment, the best performing model was Lasso. Its variable importance was measured by the final coefficients derived by the best lambda value of the shrunk variables. Only the most important variables were used for the prediction task. Figure 12 displays Lasso variable importance with absolute values of the final model

coefficients. It is seen that the majority of the importance was given to the monthly interactions, followed by a relative monthly change in arrivals (A_{rc}). Much smaller importance was given to the monthly sentiment variables (S_t , S_{t+1} , S_{t+2}), the maximum (S_{max}) and the slope direction (S_{dir}). All together the sentiment variables were given the least importance by Lasso with the coefficient values as low as 0.01 or even shrunk to a zero.

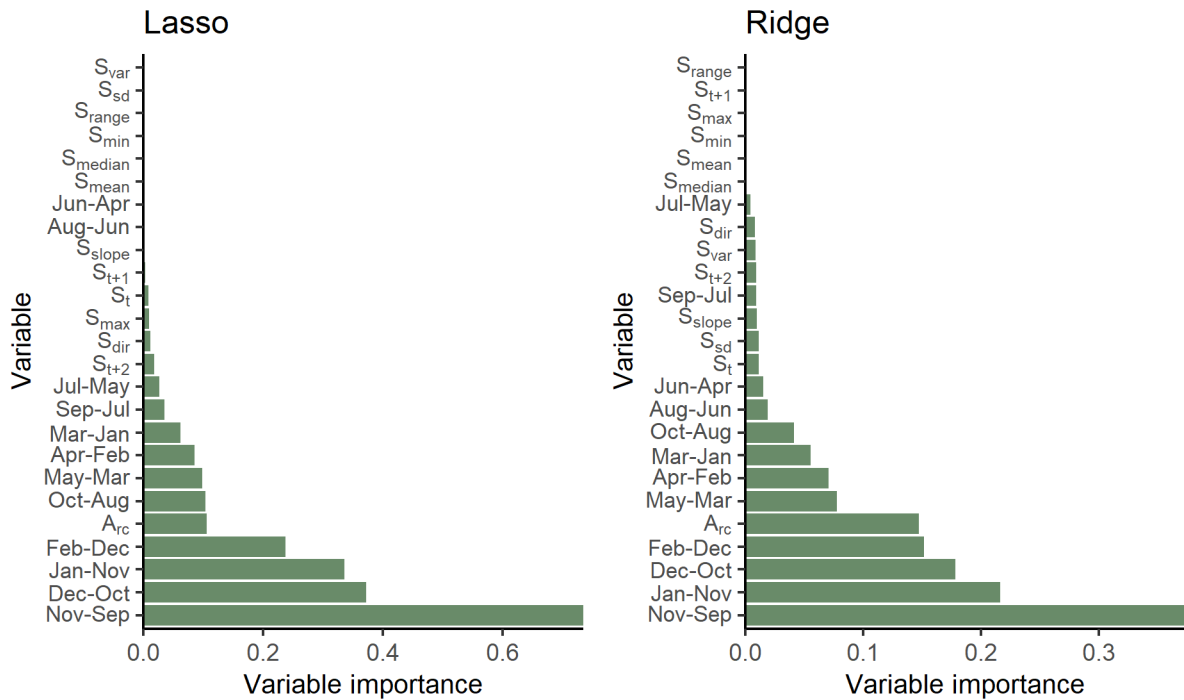


Figure 12. Lasso and Ridge variable importance.

In the second part of research question 1, the best performing model was Ridge. Analogously to Lasso, it measures the variable importance in a very similar way (except reducing values to be equal to a zero). Considering how both models measure feature importance, it was not surprising to see that the greatest importance was given to the same set of variables (Figure 12). None of the sentiment variables were included in the top ten of the most important features. Although, Ridge cannot grant zero values to any of the variables many of the sentiment variables were very close to zero and some were given equally small coefficients (approximately 0.01)

interactions rather than sentiment since only 4 sentiment variables had a t-value bigger than 1.

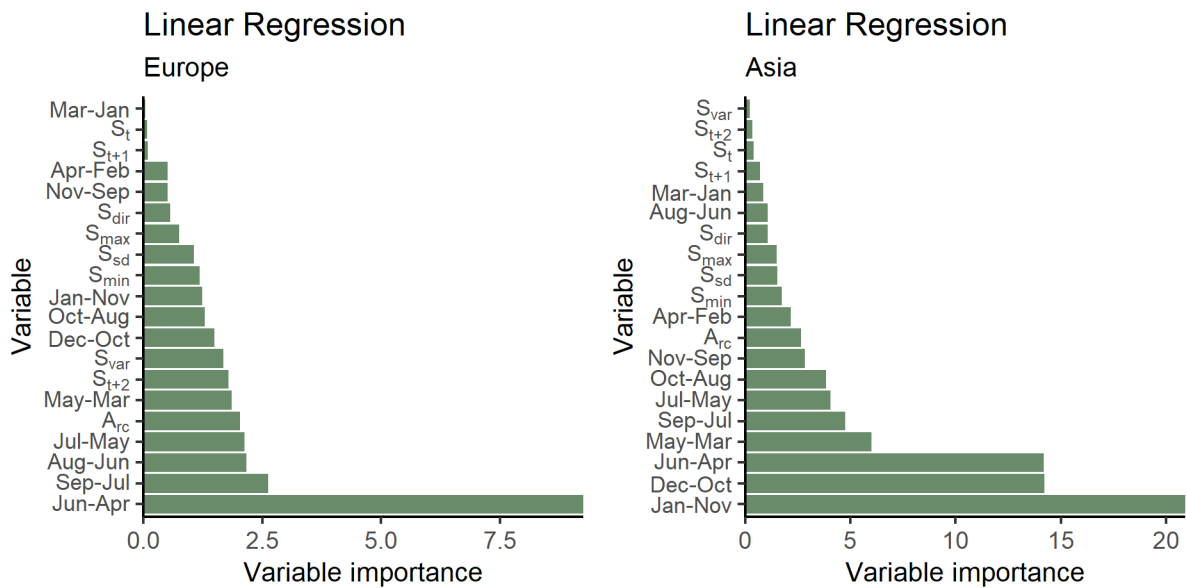


Figure 14. Linear Regression variable importance for Europe and Asia.

As a conclusion, the results from experiment 1 indicated that all the models were able to predict a relative monthly change in arrivals two months ahead relatively well, that is, all the models outperformed the baseline model. Thus, all of the predictive models were able to generalize to some extent to unseen data. However, models were not as accurate in predicting relative monthly changes in arrivals within countries in continents. The results of experiment 2 showed that the performance was highly dependent on the continent for which the predictions were being made. As a result, no single model was found to perform equally well in all cases. Therefore, the results of experiment 3 indicated that the variable importance measured in different models varied. Only the Random Forest and party Linear Regression seemed to give more importance to sentiment variables than the rest of the models.

4. DISCUSSION

4.1. REVIEW OF EXPERIMENTS

The main goal of this thesis was to predict a relative monthly change in arrivals two months ahead by incorporating two sources of data: arrivals data from the Statistics Department of New Zealand and the news sentiment extracted from the GDELT. Although the GDELT was also used by other researchers (Ahmed et al., 2016; Lin, Carley & Cheng, 2016) there is no work done with the combination of these two sources. The core dataset contained monthly intervals and included a time span of 39 years from January 1979 to December 2018 for a set of 50 countries. An average tone between New Zealand toward selected 50 countries was assumed to reflect the public sentiment (Ye, 2017). In order to preserve the time-related dependency characterizing the time series and to apply machine learning algorithms, a new data set was created. This was achieved by the sliding time window approach with a fixed window size of three months. In addition, supplementary features were included in each time window to increase the amount of data available for each regression prediction task. Afterward, the data was split into the training and testing sets. For the sake of comparison simple and more complex machine learning models were applied (Linear Regression, Ridge, Lasso, Random Forest, Support Vector Regression and Gradient Boosting Machine). Some of these models were used in other studies regarding migration prediction (Robinson & Dilkina, 2018; Ahmed et al., 2016). In order to evaluate the performance of each model, MSE was calculated for both training and testing sets. Also, a Naive model was selected as a baseline which predicted the two months ahead monthly relative change in arrivals to be the same as the previous monthly change (Hyndman & Athanasopoulos, 2018).

The goal of the research question 1 was to predict a relative monthly change in arrivals two months ahead from Australia to New Zealand based on sentiment in the time period of

April 1979 – December 2018. As a result, all six models that were tested had predictive power. However, the best performing model was Lasso with the test MSE value of 0.0156. Meaning that on average the model is 0.0156 off from actual value of a relative monthly change in arrivals. As a comparison, the Lasso test MSE was 0.078 smaller than the baseline MSE. This was not surprising since Lasso predictive power was shown in the work done by (Ahmed et al., 2016). To assess the predictive power of all the models' ability to generalize on unseen data migration flow from the United States to New Zealand in the period of July 1979 – December 2018 was used as a test set. As a result, all the models outperformed the baseline, but the best result was achieved by Ridge model. The test MSE was 0.1508 or 0.0852 less than the test MSE for the baseline model. However, an average MSE value counted for all the models (including the baseline) increased from 0.0300 for Australia data to 0.1736 for the United States data which suggested that the models cannot perform as accurate when the data comes from a different distribution with the different amplitude of changes. On the other hand, all the models were able to generalize to some extent.

The goal of the research question 2 was to find the patterns within the different continents and to determine the best predictive model. Four continents were formed: Oceania, Europe, Asia and Americas for a time period of January 2008 – December 2018. One country from each continent was selected as a test set. Analogously to research question 1, the same algorithms and the baseline method were selected. As a result, all the models outperformed the baseline. However, there was no single model that performed the best in each case scenario. The predictability of the Random Forest has shown to be the most powerful in Oceania and Americas where the training set size was 387 observations in each case. Linear Regression performed the best in Europe and Asia where the training set size was larger (1290 observations). This suggested that as more countries are included in the training sample it became harder to find the underlying patterns within various countries in continents. In

addition, a huge baseline MSE value for Europe suggested that the countries in this continent are very different in term of arrivals to New Zealand and the amplitude of changes in the test set was the widest. Moreover, the countries in Asia continent seemed to be the most similar. Even with the 10 different countries included in a training set, MSEs were relatively low (average MSE for all the models on a train set was 0.6406 and a test set - 0.7774) as oppose to MSEs in Europe's continent (5.8112 and 6.5668, respectively). This suggested that there might be a general trend of arrivals from Asia to New Zealand.

The goal of the research question 3 was to evaluate how important is the sentiment for forecasting arrivals. For this task variable importance was extracted only from these models that performed the best in the previous experiments (Lasso, Ridge, Random Forest and Linear Regression). All the models agreed on monthly interaction being the best determinant. However, the models have been very different in terms of sentiment when it comes to importance. Sentiment features were included in the top ten regarding importance only in Random Forest and Linear Regression (in Europe's case only). This suggested a relatively small impact which was not in accordance with expectations. It was demonstrated in the literature that the sentiment affects migration flows (Lumpe, 2018; Gorinas & Pytlikova, 2015). On the other hand, this was just an initial step trying to examine the importance of the sentiment in terms of migration flow. The exact sentiment importance cannot be concluded only from these models as a better methodology is needed.

4.2. LIMITATIONS AND FUTURE WORK

The research was based on selected 50 countries however more countries could be included to extend the training and testing possibilities. This could serve as a better estimation of predictive models' ability. Moreover, the research could concentrate on specific actions that perhaps could provide even better predictive power. Some examples of events coded by CAMEO are 014 – “Consider policy option”, 0211 – “Appeal for economic

cooperation“, 030 – “Express intent to cooperate”, and so on (Schrodt, 2012). Another possibility would be to extract theme-related information from GDELT Global Knowledge Graph (GKG) as was implemented by (Ahmed et al., 2016) who filtered events covering specific terms and matched these with the locations. Also, the number of events or their change could be included as a feature too (Jalal-Kamali & Pynadath, 2016). Besides that, each additional feature could be inspected in terms of prediction accuracy.

Despite possibilities provided by the GDELT, there are critics about using this sort of data. Starting with an average tone abnormality from around 2012 and lack of official documentation regarding any changes raises a concern of accuracy (Weller & McCubbins, 2014). Besides that, the completeness and correctness of articles cannot be trusted since they are not being verified and duplicates may appear. Moreover, over 30 years of existence, the volume of articles in GDELT grew exponentially and this makes comparisons across the time very vague (Ulfelder, 2014).

More machine learning algorithms could be chosen as a variety of different ones were used in previous researches. For instance, an artificial neural network model which was used to predict people movements globally and within counties in the USA (Robinson & Dilkina, 2018).

The importance of variables was evaluated only on a training set for the best performing models. However, it does not mean that the same importance is held in the test sets. Random Forest variable importance is possibly biased as a feature which is even not related to the outcome but has more splitting possibilities possess a higher chance in reducing impurity (Nembrini, Konig & Wright, 2018). Furthermore, the Random Forests variable importance is not reliable with differently scaled and varying types of variables (Strobl, Boulesteix, Zeileis & Hothorn, 2007). Both binary and numerical variables were included in the research, as well as z-score was used for the sentiment data but relative monthly

differences were calculated on absolute values. The same applies to Linear Regression feature importance – variables have to be uniformly distributed on the same range (Greenwell, Boehmke & McCarthy, 2018). Lasso shrinks highly correlated variables, which could be the reason for not including sentiment variables in the model. Moreover, it also requires the standardization of variables (Tibshirani, 1997). In order to evaluate feature importance, more robust measures should be applied specifically designed for that purpose to achieve accurate conclusions. Besides that, the experiment should be run multiple times at each time including a different set of variables and measuring an error rate. Only then the importance of sentiment variable in arrivals prediction can be concluded.

For the sliding time window procedure, a window of the fixed size of three months was selected. However, there is no guarantee of this number being the most suitable in predicting migration flow. As noted by researchers (Nakajima, Matsui, Sakai & Yagi, 2011) a small window size has a much larger variation. In case of extraordinary observations, for instance, an extreme decrease, an enduring impact of data point being predicted will be observed within that window size. After all, the differences were observed between the continents regarding extraordinary scale. This could be one of the reasons why some models performed worse than the others. The more advanced method would be trying different window sizes, evaluating each and only then selecting the best one.

To conclude, this thesis has focused on predicting migration flow to New Zealand. Under the assumption that the news sentiment between the two countries can improve predictive accuracy, a new dataset from migration flows and sentiment indicators was constructed. Machine learning algorithms achieved promising results. However, due to the lack of previous literature regarding both of these specific datasets it is not possible to make any comparisons. More research is needed in order to enable the full potential of the methods used.

5. CONCLUSION

With regards to the complexity of migration determinants, imperfect theory and possibly biased forecasts for migration prediction, this thesis explored and developed a possible approach for migration prediction. Focusing on New Zealand and assuming that the news sentiment between the two countries can improve predictive accuracy, a new dataset involving migration flow and the news sentiment was constructed. The Statistics Department of New Zealand was used to extract the number of arrivals and the GDELT database was used to extract the news sentiment. Additionally, the sliding time window approach suitable for time-series data was used to construct the temporal features. Based on this, supervised machine learning algorithms were developed to predict monthly relative changes in arrivals to New Zealand two months ahead. The key findings of this thesis are:

Key finding 1: The machine learning models compared with the Naïve model provide the potential for improving the migration prediction accuracy at a country level and have the ability to generalize to some extent to unseen data.

Key finding 2: The models are not as accurate in predicting relative monthly changes in arrivals within countries in continents. Thus, results are highly dependent on the continent being used for prediction. In addition to that, no single model is found to perform equally well in all experiments.

Key finding 3: The importance of the sentiment in various predictive models varies. It is found to have a minimal impact in Lasso and Ridge models while in Random Forest and partly Linear Regression the influence is greater.

Key finding 4: The methodology used in this thesis can serve as a starting point for building the migration prediction models and can give a brief examination of determinants affecting migration flow. However, improving the feasibility of the technique remains one of the main directions in the future.

References

- Ahmed, M. N., Barlacchi, G., Braghin, S., Calabrese, F., Ferretti, M., Lonij, V., Nair, R., Novack, R., Paraszczak, J., & Toor, A. S. (2016). A Multi-Scale Approach to Data-Driven Mass Migration Analysis. Retrieved from http://ceur-ws.org/Vol-1831/paper_4.pdf
- Arel-Bundock, V., Enevoldsen, N., & Yetman, C. (2018). Countrycode: An R package to convert country names and country codes. *Journal of Open Source Software*, 3(28), 848. doi:10.21105/joss.00848
- Azose, J. J., & Raftery, A. E. (2015). Bayesian Probabilistic Projection of International Migration. *Demography*, 52(5), 1627-1650. doi:10.1007/s13524-015-0415-0
- Bijak, J., Disney, G., Findlay, A. M., Forster, J.J., Smith, P. W. F., & Wisniowski, A. (2019). Assessing time series models for forecasting international migration: Lessons from the United Kingdom. *Journal of Forecasting*. DOI: 10.1002/for.2576
- Bijak, J., Disney, G., Wisniowski, A., Forster, J., Smith, P., & Findlay, A. (2015). How to forecast international migration. *ESRC Centre for Population Change. Briefing 28*. Retrieved from http://www.cpc.ac.uk/docs/BP28_How%20to%20forecast%20international%20migration.pdf
- Boehmke, B. (2018, June 14). Gradient Boosting Machines [Blog post]. Retrieved from <http://uc-r.github.io/2018/06/14/gbm-regression/>
- Bohme, M., Groger, A., & Stohr, T. (2018). Searching for a Better Life: Predicting International Migration with online Search Keyword. Retrieved from https://www.ifw-kiel.de/fileadmin/Dateiverwaltung/IfW-Publications/Tobias_Stoehr/searching-for-a-better-life-predicting-international-migration-with-online-search-keywords/searching-for-a-better-life-predicting-international-migration-with-online-search-keywords.pdf
- Bontempi, G., & Taieb, S. B. (2011). Conditionally dependent strategies for multiple-step-ahead prediction in local learning. *International Journal of Forecasting*, 27(3), 689-699. doi:10.1016/j.ijforecast.2010.09.004
- Boyle, B. H. (2011). *Support vector machines: Data analysis, machine learning, and applications*. New York, NY: Nova Science.
- Buettner, T., & Muenz, R. (2016). Comparative Analysis of International Migration in Population Projections. *Knomad Working Paper 10*. Retrieved from <https://www.knomad.org/sites/default/datasets/2017-04/KNOMAD%20WP%2010%20Comparative%20Analysis%20of%20International%20Migration%20in%20Population%20Projections.pdf>
- Cherkassky, V., & Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks*, 17(1), 113-126. doi:10.1016/s0893-6080(03)00169-2

- Conway, D., & White, J. M. (2012). *Machine learning for hackers*. Sebastopol, CA: O'Reilly Media.
- Crush, J., & Ramachandran, S. (2010). Xenophobia, International Migration and Development. *Journal of Human Development and Capabilities*, 11(2), 209-228. doi:10.1080/19452821003677327
- Delias, P., Zoumpoulidis, V., & Kazanidis, I. (2019). Visualizing and exploring event databases: A methodology to benefit from process analytics. *Operational Research*. doi:10.1007/s12351-018-00447-z
- Disney, G., Wisniowski, A., Forster, J.J., Smith P. W. F., Bijak, J. (2015). Evaluation of existing migration forecasting methods and models. Retrieved from https://www.research.manchester.ac.uk/portal/datasets/32468230/FULL_TEXT.PDF
- Elshendy, M., & Colladon, A. F. (2017). Big data analysis of economic news: Hints to forecast macroeconomic indicators. *International Journal of Engineering Business Management*, 9, 1-12. doi:10.1177/1847979017720040
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1-22. Retrieved from <http://www.jstatsoft.org/v33/i01/>.
- Galloppo, G. (2010). Sample period selection with variable rolling windows: Searching for beta stability. Retrieved from <https://www.aabri.com/LV2010Manuscripts/LV10118.pdf>.
- Gorinas, C., & Pytliková, M. (2015). The Influence of Attitudes toward Immigrants on International Migration. *International Migration Review*, 51(2), 416-451. doi:10.1111/imre.12232
- Greenwell, B. M., Boehmke, B. C., & McCarthy, A. J. (2018). A Simple and Effective Model-Based Variable Importance Measure. Retrieved from <https://arxiv.org/pdf/1805.04755.pdf>.
- Greenwell, B., Boehmke, B., Cunningham, J., & GBM Developers (2019). gbm: Generalized Boosted Regression Models. [R package version 2.1.5]. Retrieved from <https://CRAN.R-project.org/package=gbm>
- Grolemund, G., & Wickham, H. (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software*, 40(3), 1-25. Retrieved from <http://www.jstatsoft.org/v40/i03/>.
- Haptonstahl, S. R., Scherer T., Thoms, T., & Wheatley, P. (2014). GDELTtools: Download, slice, and normalize GDELT data. [R package version 1.2]. Retrieved from <https://CRAN.R-project.org/package=GDELTtools>
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.

- Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.) [2]. Retrieved May 7, 2019, from <https://otexts.com/fpp2/>
- Inoue, A., Jin, L., & Rossi, B. (2017). Rolling window selection for out-of-sample forecasting with time-varying parameters. *Journal of Econometrics*, 196(1), 55-67.
- International Travel and Migration. Statistics New Zealand. (2018, March 14). Retrieved from http://datainfoplus.stats.govt.nz/Item/nz.govt.stats/f705ca38-ea6e-453f-b1d9-a95dc0fc59?_ga=2.108461058.1674334919.155048383678449077.1550482834#nz.govt.stats/d83affa2-c0e1-4629-9aff-cc8452eb41c6#
- Jalal-Kamali, A., & Pynadath, D. V. (2016). Toward a Bayesian Network Model of Events in International Relations. *Social, Cultural, and Behavioral Modeling Lecture Notes in Computer Science*, 311-322. doi:10.1007/978-3-319-39931-7_30
- James, G., Witten, D., Hastie, T. J., & Tibshirani, R. J. (2017). *An introduction to statistical learning: With applications in R*. New York, NY: Springer.
- Kuhn, M. (2018). Caret: Classification and Regression Training. [R package version 6.0-81]. Retrieved from <https://CRAN.R-project.org/package=caret>
- Kumar, S., Benigni, M., & Carley, K. M. (2016). The impact of US cyber policies on cyber-attacks trend. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on* (pp. 181-186). IEEE. Retrieved from: <http://www.casos.cs.cmu.edu/publications/papers/2016ImpactofUSCyber.pdf>
- Lahiri, S. K., & Ghanta, K. C. (2008). The Support Vector Regression with the parameter tuning assisted by a differential evolution technique: Study of the critical velocity of a slurry flow in a pipeline. *Chemical Industry and Chemical Engineering Quarterly*, 14(3), 191-203. doi:10.2298/ciceq0803191l
- Leetaru, K., Schrodtt, P. A. (2013). GDELT: Global Data on Events, Location and Tone, 1979-2012. Retrieved from <http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf>
- Liaw, A., & Wiener, M. (2002). Classification and Regression by RandomForest. *R News*, 2/3, 18-22. Retrieved May 7, 2019, from https://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf.
- Lin, L., Carley, K. M., & Cheng, S. (2016). An agent-based approach to human migration movement. *2016 Winter Simulation Conference (WSC)*. doi:10.1109/wsc.2016.7822380
- Lumpe, C. (2018). Public beliefs in social mobility and high-skilled migration. *Journal of Population Economics*, 32(3), 981-1008. doi:10.1007/s00148-018-0708-x
- Meseguer, C., & Kemmerling, A. (2018). What Do You Fear? Anti-immigrant Sentiment in Latin America. *International Migration Review*, 52(1), 236-272. doi:10.1111/imre.12269

- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2019). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. [R package version 1.7-1]. Retrieved from <https://CRAN.R-project.org/package=e1071>
- Nakajima, T., Matsui, T., Sakai, Y., & Yagi, N. (2011). Capturing Changes in Asymmetric Price Transmission: A Rolling Window TAR Estimation Using Bluefin Tuna Case Study. *Journal of International Fisheries*, 10. Retrieved from <http://jifrs.info/Journal/Vol.10,Nakajima.pdf>
- Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7. doi:10.3389/fnbot.2013.00021
- Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the Gini importance? *Bioinformatics*, 34(21), 3711-3718. doi:10.1093/bioinformatics/bty373
- New Zealand Immigration Statistics (n.d.). Arrivals and Departures. Retrieved from <http://www.migrationstats.com/nz/ArrivalDeparture>
- Raymer, J., & Wiśniowski, A. (2018). Applying and testing a forecasting model for age and sex patterns of immigration and emigration. *Population Studies*, 72(3), 339-355. doi:10.1080/00324728.2018.1469784
- Robinson, C., & Dilkina, B. (2018). A Machine Learning Approach to Modeling Human Migration. *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) - COMPASS 18*. doi:10.1145/3209811.3209868
- Rojas-Romagosa, H., Bollen, J. (2018). Estimating migration changes from the EU's free movement of people principle. *CPB Discussion Paper*. Retrieved from <https://www.cpb.nl/sites/default/datasets/omnidownload/CPB-Discussion-Paper-385-estimating-migration-changes-from-the-EUs-free-movement-of-people-principle.pdf>
- RStudio Team (2016). RStudio: Integrated development for R [Computer software]. RStudio, Inc., Boston, MA. Retrieved from: <http://www.rstudio.com/>
- Schrodt, P. A. (2012). CAMEO Conflict and Mediation Event Observations Event and Actor Codebook. Retrieved from <http://data.gdeltpproject.org/documentation/CAMEO.Manual.1.1b3.pdf>
- Scornet, E. (2018). Tuning Parameters in Random Forests. *ESAIM: Proceedings And Surveys*, 60, 144-162. Retrieved May 7, 2019, from <https://www.esaim-proc.org/articles/proc/pdf/2017/05/proc186008.pdf>.
- Sherkat, D. E., & Lehman, D. (2018). Bad Samaritans: Religion and Anti-Immigrant and Anti-Muslim Sentiment in the United States. *Social Science Quarterly*, 99(5), 1791-1804. doi:10.1111/ssqu.12535
- Statistics New Zealand. (n.d.). *International Travel and Migration – ITM: Permanent & long-term migration by every country of residence and citizenship (Monthly)*. [Data file]. Available from <http://archive.stats.govt.nz/infoshare/Default.aspx>

- Strobl, C., Boulesteix, A., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1). doi:10.1186/1471-2105-8-25\
- The GDELT Project. (n.d.). *GDELT 2.0 Event database*. [Data file]. Available from <https://www.gdeltproject.org/>
- The GDELT Project (2015). The GDELT Event Database data format codebook V2.0. Retrieved from http://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf
- Tibshirani, R. (1997). The Lasso Method For Variable Selection In The Cox Model. *Statistics in Medicine*, 16(4), 385-395. doi:10.1002/(sici)1097-0258(19970228)16:43.0.co;2-3
- Ulfelder, J. (2014, June 06). Another Note on the Limitations of Event Data [Blog post]. Retrieved from <https://dartthrowingchimp.wordpress.com/2014/06/06/another-note-on-the-limitations-of-event-data/>
- Weller, N., & McCubbins, K. (February 20, 2014). Raining on the parade: Some cautions regarding the Global Database of Events, Language and Tone dataset [Blog post]. Retrieved from <https://politicalviolenceataglance.org/2014/02/20/raining-on-the-parade-some-cautionsregarding->
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. Retrieved from <https://ggplot2.tidyverse.org>
- Wickham, H., & Henry, L. (2018). *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. [R package version 0.8.2]. Retrieved from: <https://CRAN.R-project.org/package=tidyr>
- Wickham, H., François, R., Henry, L., & Müller, K. (2018). *dplyr: A Grammar of Data Manipulation*. [R package version 0.7.8]. Retrieved from: <https://CRAN.R-project.org/package=dplyr>
- Witten, I. H., & Frank, E. (2009). *Data mining: Practical machine learning tools and techniques with Java implementations* (2nd ed.). San Diego, CA: Elsevier Science & Technology Books.
- Wladyka, D. K. (2017). Queries to Google Search as Predictors of Migration Flows from Latin America to Spain. *Journal of Population and Social Studies*, 25(4), 312 – 327. DOI: 10.25133/JPSSv25n4.002
- Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1), 1-17. doi:10.18637/jss.v077.i01
- Ye, M. (2017). *China-South Korea relations in the new era: Challenges and opportunities*. Lanham, MD: Lexington Books.

Zagheni, E., Garimella, V. R., Weber, I., & State, B. (2014). Inferring international and internal migration patterns from Twitter data. *Proceedings of the 23rd International Conference on World Wide Web - WWW 14 Companion*.
doi:10.1145/2567948.2576930

Appendix A: List of countries

Table 1.
Selected countries.

<i>Nr</i>	<i>Country Name</i>	<i>Country Code</i>	<i>Nr</i>	<i>Country Name</i>	<i>Country Code</i>
1	Australia	AUS	26	Italy	ITA
2	China	CHN	27	Sweden	SWE
3	United Kingdom	GBR	28	Spain	ESP
4	United States of America	USA	29	Denmark	DNK
5	Germany	DEU	30	Argentina	ARG
6	Japan	JPN	31	Chile	CHL
7	India	IND	32	Austria	AUT
8	France	FRA	33	Belgium	BEL
9	Korea	KOR	34	Vietnam	VNM
10	Canada	CAN	35	Czech Republic	CZE
11	Malaysia	MYS	36	Sri Lanka	LKA
12	Fiji	FJI	37	Poland	POL
13	Netherlands	NLD	38	Israel	ISR
14	South Africa	ZAF	39	Finland	FIN
15	Philippines	PHL	40	Saudi Arabia	SAU
16	Taiwan	TWN	41	Norway	NOR
17	Singapore	SGP	42	Mexico	MEX
18	Tonga	TON	43	Vanuatu	VUT
19	Ireland	IRL	44	Pakistan	PAK
20	Western Samoa	WSM	45	Ukraine	UKR
21	Hong Kong	HGK	46	Portugal	PRT
22	Thailand	THA	47	Colombia	COL
23	Indonesia	IDN	48	Hungary	HUN
24	Brazil	BRA	49	Cambodia	KHM
25	Switzerland	CHE	50	Iran	IRN

Note. Here these countries represent the country of origin for arrivals and the ‘Actor2’ in the GDELT Event data. New Zealand is not included as it is a country of destination for arrivals and it was selected as an ‘Actor1’ in the GDELT Event data.

After missing values identification only countries in bold text were selected for further research (in total 32).

Appendix B: Missing months per year

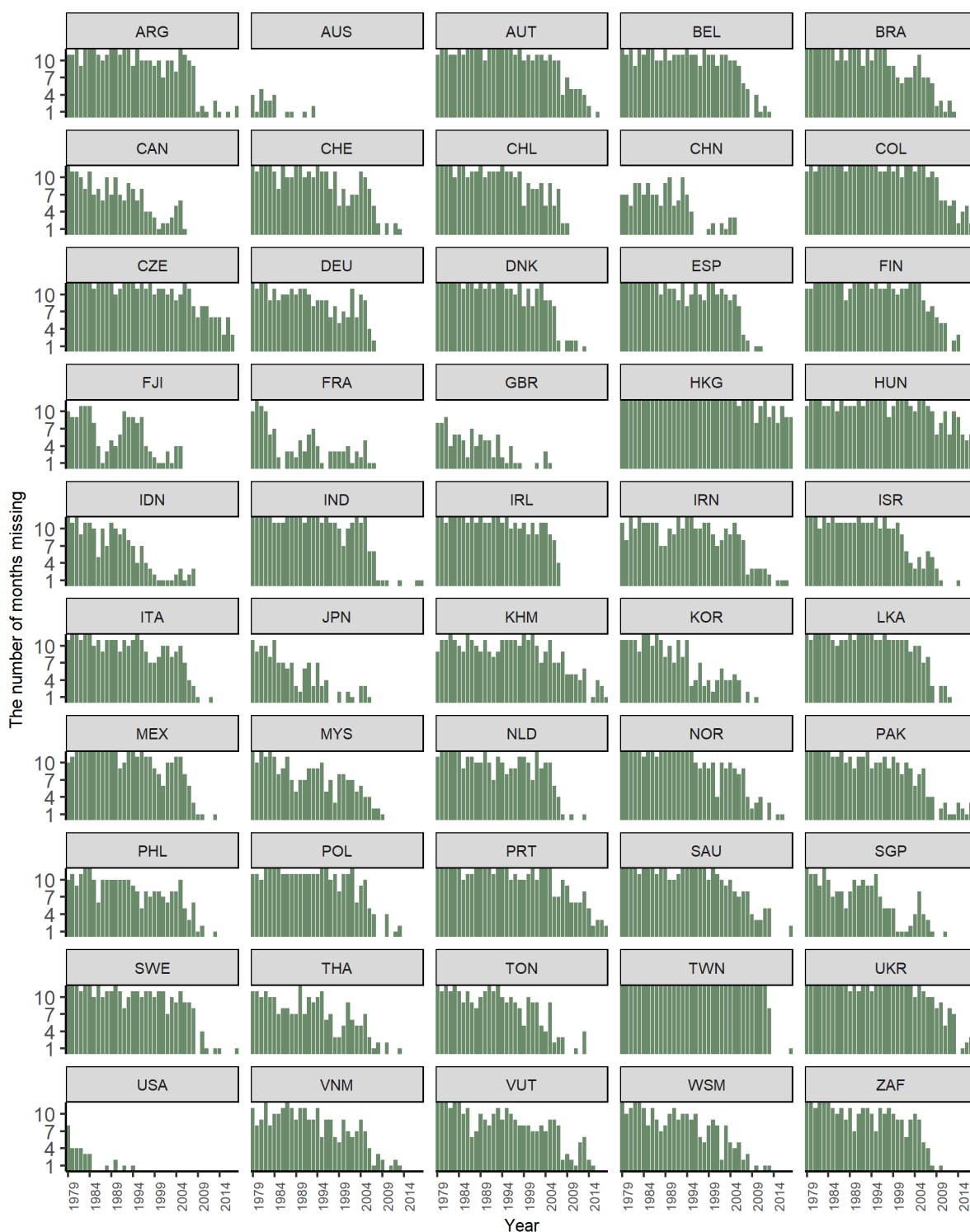


Figure 1. *The number of missing months per year in each country.*

Appendix C: Average tone before and after z-scores for Australia and the United States

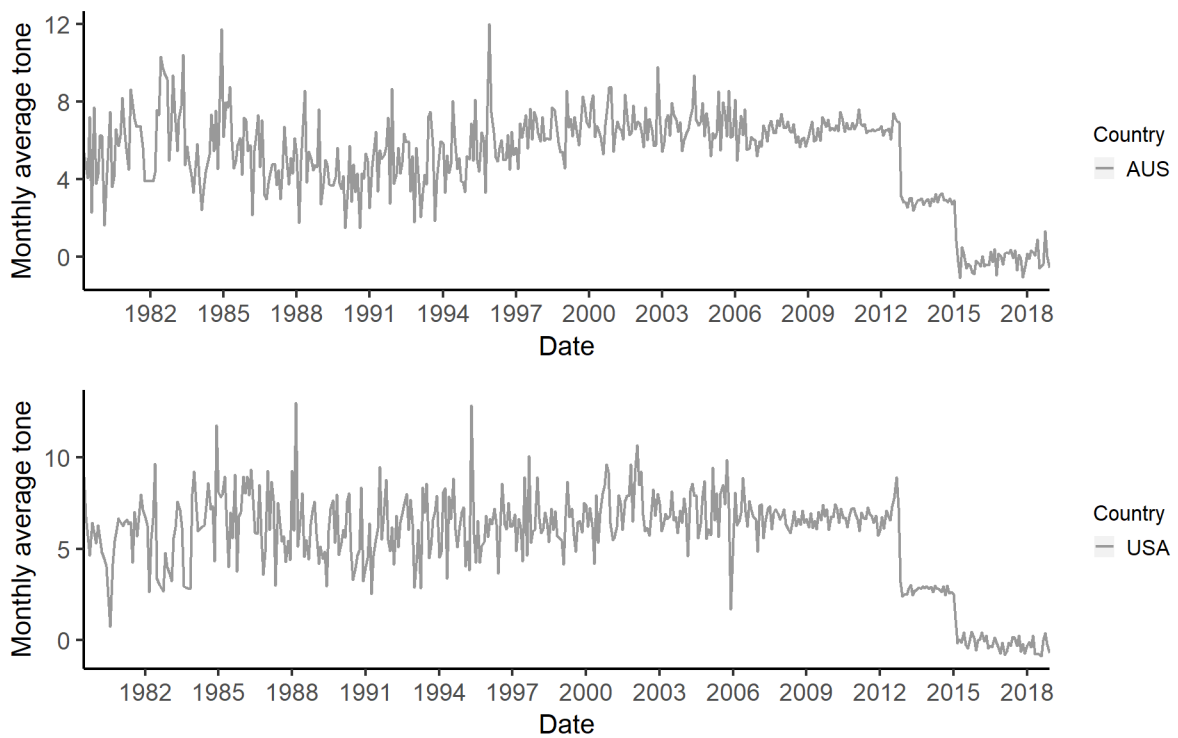


Figure 1. Average tone before z-scores implementation.

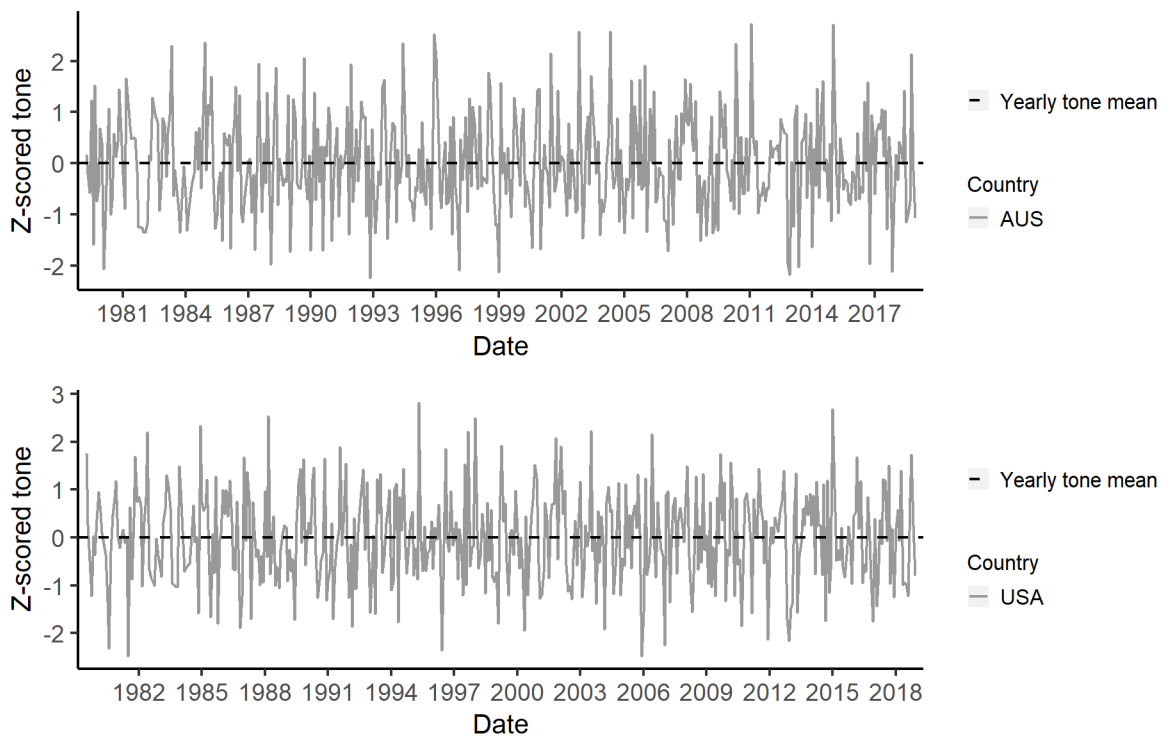


Figure 2. Average tone after z-scores implementation.

Appendix D: Seasonality detection

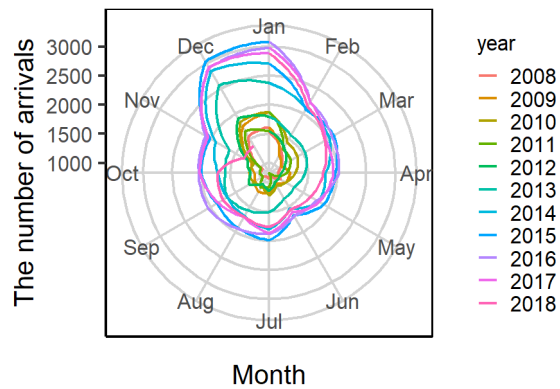


Figure 1. Monthly polar seasonal plot for all the countries per year.

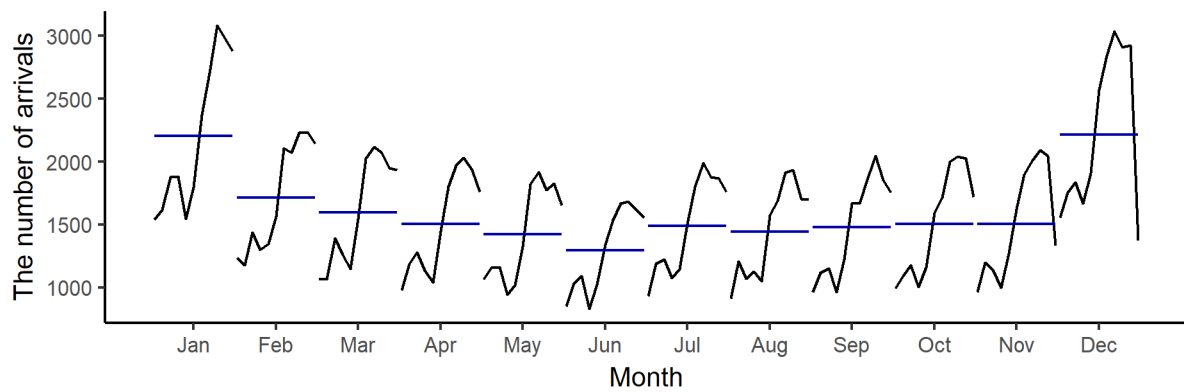


Figure 2. Monthly arrivals average for all the countries in 2008-2018.

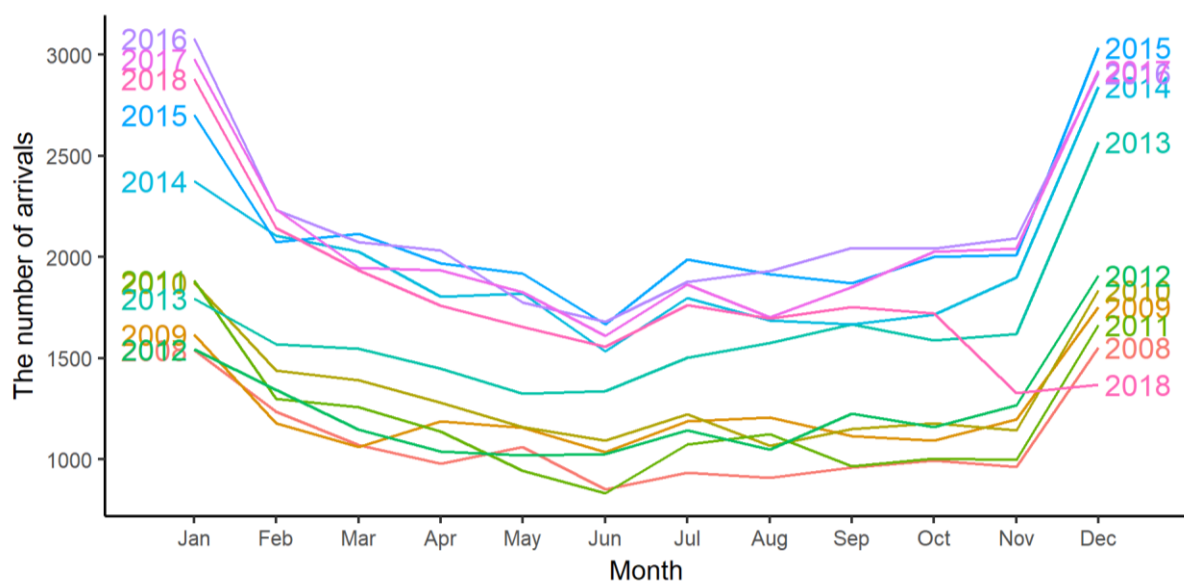


Figure 3. Monthly seasonal plot for all the countries per year.

Appendix E: Libraries

R Studio (version 3.5.1) was used to load, prepare, analyze the data, tune and run different algorithms (RStudio Team, 2016). To load GDELT dataset ‘GDELTtools’ package was used (Haptonstahl, Scherer, Thoms, & Wheatley, 2014). For data cleaning and preparation ‘tidyr’ (Wickham & Henry, 2018), ‘dplyr’ (Wickham, François, Henry & Müller, 2018) and ‘lubridate’ (Grolemund & Wickham, 2011) packages were used and for results visualizations - ‘ggplot2’ (Wickham, 2016). To convert country names to the country codes ‘countrycode’ package was used (Arel-Bundock, Enevoldsen & Yetman, 2018).

Packages used for running the algorithms: ‘caret’ (Kuhn, 2018), ‘glmnet’ (Friedman, Hastie & Tibshirani, 2010), ‘randomForest’ (Liaw & Wiener, 2002), ‘ranger’ (Wright & Ziegler, 2017), e1071 (Meyer, Dimitriadou, Hornik, Weingessel & Leisch, 2019), ‘gbm’ (Greenwell, Boehmke, Cunningham, & GBM Developers, 2019).

Appendix F: Results of experiment 1

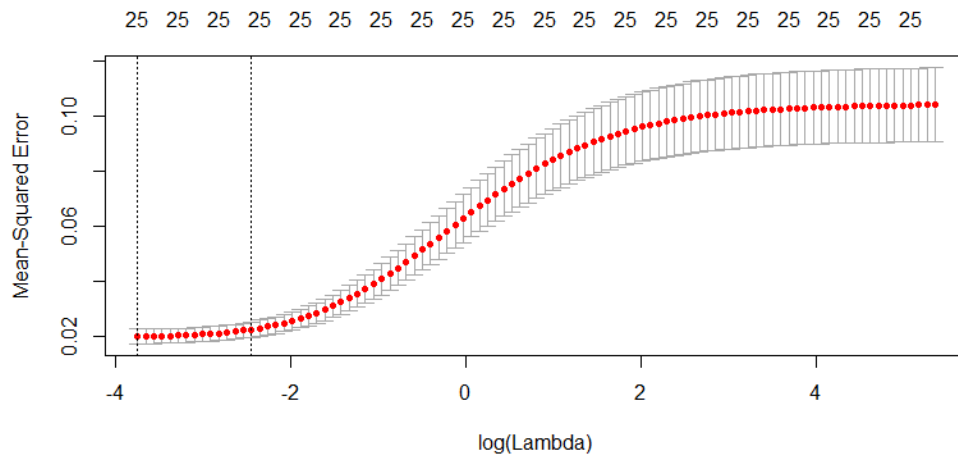


Figure 1. *Lambda selection for Ridge.*

Note: The minimum lambda value is 0.02351.

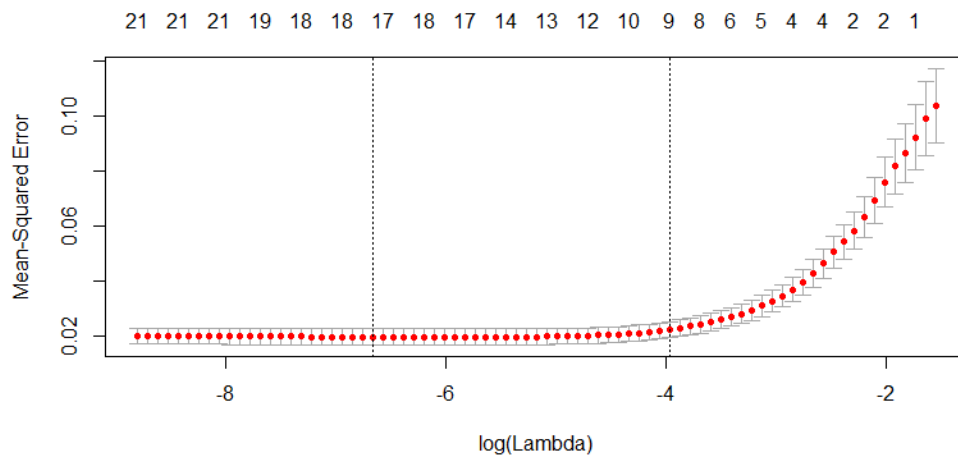


Figure 2. *Lambda selection for Lasso.*

Note: The minimum lambda value is 0.0013.

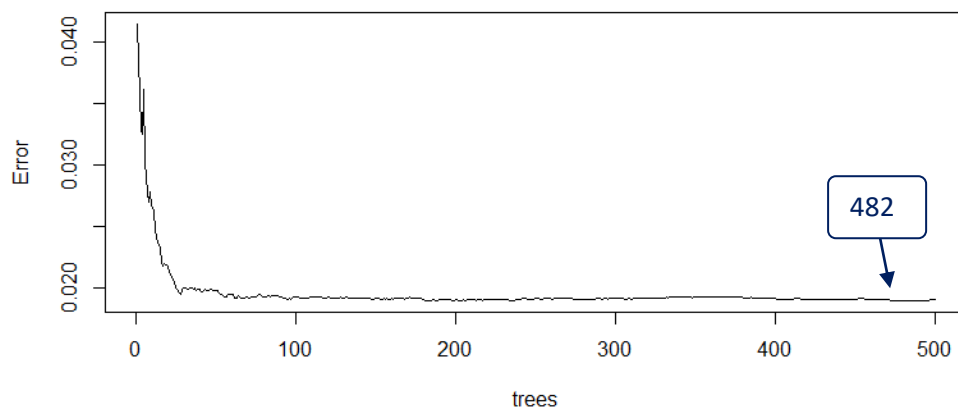


Figure 3. *Optimal number of trees for Random Forest.*

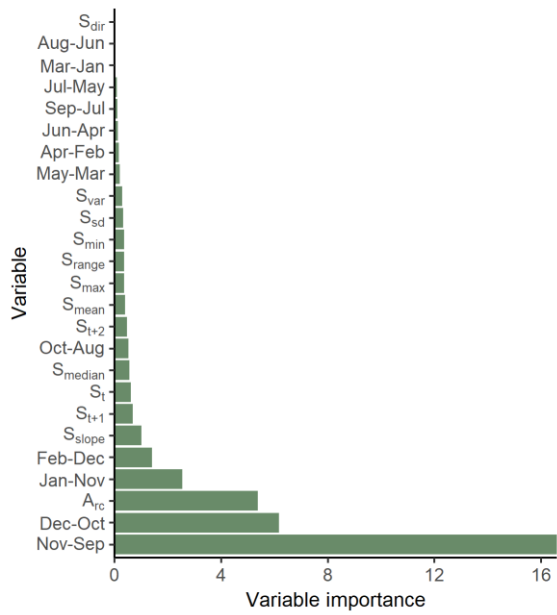


Figure 4. Variable importance for Random Forest.

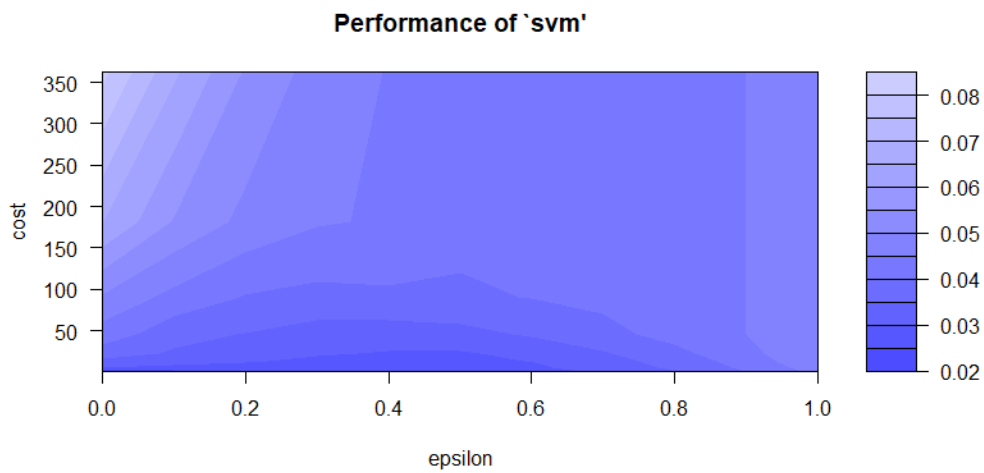


Figure 5. Performance of SVM.

Note. The best performance is achieved with epsilon value equal to 0 and cost – 1.4142.

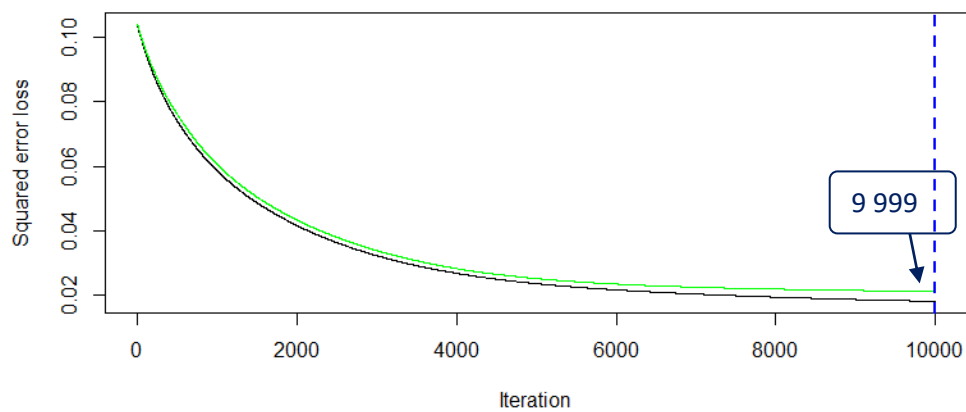


Figure 6. GBM optimal number of trees.

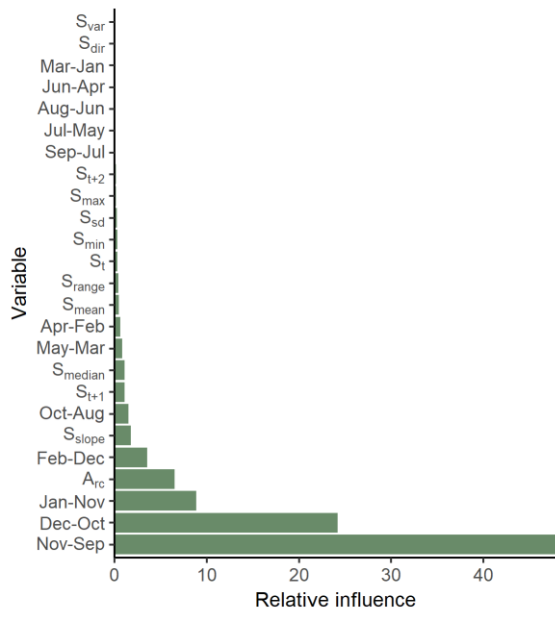


Figure 7. *GBM variable importance.*

Appendix G: Results of experiment 3

Table 1.

Relative variable importance from the best performing models in research question 1 and 2.

	<u>Australia</u>	<u>United States</u>	<u>Europe</u>	<u>Asia</u>	<u>Oceania</u>	<u>Americas</u>
	<i>Lasso</i>	<i>Ridge</i>	<i>Linear Regression</i>	<i>Linear Regression</i>	<i>Random Forest</i>	<i>Random Forest</i>
S_t	0.0088	0.0118	0.0789	0.4195	1.8060	16.0041
S_{t+1}	0.0038	0.0005	0.0969	0.7101	2.0268	8.4140
S_{t+2}	0.0190	0.0092	1.7961	0.3583	1.3312	15.7488
S_{mean}	0.0000	0.0011	NA	NA	0.9168	7.9962
S_{min}	0.0000	0.0009	1.1913	1.7407	1.0370	11.4759
S_{max}	0.0101	0.0006	0.7558	1.4871	1.1417	12.3048
S_{median}	0.0000	0.0013	NA	NA	1.1762	5.6695
S_{sd}	0.0000	0.0116	1.0684	1.5351	0.9380	15.8218
S_{var}	0.0000	0.0086	1.6882	0.2241	1.0504	12.4085
S_{slope}	0.0000	0.0100	NA	NA	1.5755	24.2899
S_{dir}	0.0123	0.0082	0.5705	1.0862	0.0360	0.2535
S_{range}	0.0000	0.0002	NA	NA	0.9729	12.5143
A_{rc}	0.1058	0.1474	2.0399	2.6599	3.6062	22.7287
<i>MarJan</i>	0.0623	0.0554	0.0409	0.8491	0.9739	2.3268
<i>AprFeb</i>	0.0854	0.0707	0.5155	2.1780	0.0378	0.2751
<i>MayMar</i>	0.0989	0.0781	1.8559	6.0009	0.0904	1.9973
<i>JunApr</i>	0.0000	0.0153	9.2387	14.1710	0.1272	1.1804
<i>JulMay</i>	0.0275	0.0043	2.1311	4.0463	0.4316	30.5625
<i>AugJun</i>	0.0000	0.0191	2.1717	1.0622	0.0997	9.7310
<i>SepJul</i>	0.0358	0.0096	2.6327	4.7632	0.0849	1.7988
<i>OctAug</i>	0.1046	0.0416	1.2969	3.8296	0.4400	0.7695
<i>NovSep</i>	0.7348	0.3731	0.5189	2.8487	11.2993	0.3745
<i>DecOct</i>	0.3728	0.1783	1.4998	14.2252	3.7572	0.2900
<i>JanNov</i>	0.3364	0.2165	1.2385	20.8969	0.1181	0.7461
<i>FebDec</i>	0.2376	0.1518	NA	NA	0.2943	5.3945