New Zealand Domestic Net Migration: An investigation of the Environmental drivers moving people around New Zealand

DATA 422 2022S2 Group Project Report

Team JIFS: Josie Stockill – Indira Jinadasa – Frazer French – Samuel Love

Overview and Project Aim

This dataset was created with the aim of addressing the reasons people move around New Zealand, and whether there is a relation between the changing natural landscape and where people choose to reside. The final product of this project is an exploratory dataset that researchers can use to investigate what factors are causing migration in each region.

This was motivated by the changing of New Zealand's climate to determine whether people are moving due to these effects. In understanding these effects, planners may be able to attract more people to their region by addressing the underlying drivers of movement and provide assurance on regional plans to deal with these issues.

The data consists of two separate data-frames.

The historical measurements are for earthquakes, internal migration, and weather (rainfall, temperature, and wind gusts), covering the years 2014-17 (inclusive).

The second dataset is for searise in these regions covering predictions for a range of scenarios out to 2300. The prediction to be examined can be changed via a function, which allows the user to investigate what time frame for searise people are concerned with and in what region.

Earthquakes - Josie

Data sources

This analysis used data exported from Geonet's Quake Search[1] feature. Geonet is a partnership between a number of government agencies and was established in 2001 to provide monitoring of geological hazard information for New Zealand. The second dataset used in the analysis was a shapefile of regional geographic boundaries, provided by StatisticsNZ (via the koordinates [2]website).

Reasons for the data sources

Earthquakes are not unique to New Zealand but they are an environmental factor that may influence people's decisions about whether they would choose to live in New Zealand cities. Even within New Zealand, there are varying frequencies and intensities of earthquakes across different regions. It is hard to ignore the devasting potential of these natural disasters while studying through the University of Canterbury based in Christchurch. As we know, Christchurch

[1] https://quakesearch.geonet.org.nz/

[2] https://koordinates.com/from/datafinder.stats.govt.nz/layer/106666/download/

was severely affected by a major earthquake in 2011 that caused severe damage and loss of life. We wanted to enable researchers to ask the questions:

- Which regions in New Zealand experience a higher frequency of earthquakes?

- How might earthquake frequency and/or magnitude affect potential residents' decision to move to a region to live, taken into consideration alongside other environmental factors?

Difficulties in the data wrangling process

The most difficult part of wrangling this data was the import of the shapefile. This required research into the best R packages and techniques to be able to aggregate earthquakes (by their lat/long coordinates) into a count per geographical region. I faced difficulty reading the file into the notebook before I understood that the shapefile was provided with co-dependent files within the zip download.

Techniques used

The sf package in R was integral in processing the multipolygon geometry shapefile. A geometric join (using st_join()) combined the earthquake data points and regional boundaries into one dataframe. There were plenty of data wrangling techniques employed such as; selecting columns, assigning substrings as new column values, converting data types of columns, filtering dataframes by criteria, grouping and summarising dataframes to aggregate data into sum() and mean() variables, and plotting to get a quick overview of the data.

Internal migration - Indira

Data sources

This part of the project has used a dataset from 'stats.nz' which contains an enormous amount of data which is related to NZ. We have exported the data from "Internal migration estimates using linked administrative data: 2014–17". 'Internal migration estimates using linked administrative data: 2014–17' shows the potential for using administrative address data in the Integrated Data Infrastructure (IDI) as a data source for measuring the movements of New Zealand residents within the country.

Reasons for the data sources

Data is produced at a very granular level, making it possible to aggregate to different geographic areas. In our case we are going to aggregate into regions so then we could join them with several environmental factors which have occurred in respected regions and get an idea about the factors that might have affected the migration details.

Difficulties in the data wrangling process

We had to face difficulties when matching the districts into regions. Since the initial dataset contained districts and we wanted to convert them into regions to merge all. And another difficulty that we faced was omitting the same district-to-district migration and same region-to-region migration data in our final data frame which had no use for our purpose.

Techniques used

There were plenty of data wrangling techniques employed such as loading a CSV, data scraping, selecting columns, assigning substrings as new column values, renaming the columns, joining data frames, filtering data frames by criteria, grouping and summarizing data frames to aggregate data into sum (), and plotting using Julia to get a quick overview of the data.

SeaRise - Frazer

Data sources

This data was retrieved from the NZ SeaRise data website, which is a NZ project aimed at providing guidance for planners and the public about sea level change. SeaRise combines the vertical land movement of NZ's coast with searise for a number of potential scenarios, to give a range of probable estimates for the water level change along the coast in NZ.

Reasons for the data sources

NZ is a coastal nation, with over two-thirds of our population living within 5 km of the coast. As such, changing sea levels will have impacts on NZ's infrastructure, individual dwellings and our places of recreation. Even at the time of creating this report, several regions in New Zealand are already experiencing the effects of increased sea level, such as several roads in Wellington during peak floods and big swells, and homes in certain areas slipping into the sea or requiring extensive protective works to preserve their structural integrity.

Potential questions to be answered, will be what searise scenarios, if any, do people look at when choosing where to live, and how far out are they looking.

Techniques used

Working with the Searise data was relatively straightforward, as no API exists to obtain the data in a meaningful format. The data was downloaded in CSV form, and is included in the project.

The complexities in dealing with the data were in generating an output for the main table. As the data doesn't have yearly statistics like the other data used in this report, the SeaRise data exists in its own notebook and can generate outputs through a customisable function depending on the selected sea level rise scenario. Although this is not a yearly projection, it could serve well when comparing between regions to nvestigate what is driving migration

Difficulties in the data wrangling process

The primary difficulty was working with the Northland and Southland regions, as they have such large areas of coast that the data is split into several files. It was decided that they would be appended to each other through a row_bind, and processed as a single dataset. This allowed averages to be generated that reflected sea level change in the entire region.

In addition to this, there were issues in dealing with the datatypes in the searise datasets. Site IDs varied in naming convention between numbersd and characters, so for some datasets, the column type was detected as integers, which caused issues when interacting with integer type names. These were recast as characters to avoid conflict.

Weather - Sam

Data sources

The weather data is sourced from the StatsNZ website as downloadable zip files (Temperature, Rainfall, Windspeed). The chosen categories of data were temperature, rainfall, and wind gusts. The zip files contained multiple csv files including daily, site, and trend datasets.

Reasons for the data sources

StatsNZ is a trusted and comprehensive source of data pertaining to NZ and operated by the government. The chosen datasets were in turn sourced from NIWA which is a Crown Research Institute. The data on StatsNZ pertains to 30 sites around New Zealand which was ideal for our purposes of examining trends in the 16 regions of New Zealand.

Difficulties in the data wrangling process

The downloaded Zip files contained many irrelevant csv files. The ones of interest were called state_data.csv and were identical in this respect across the three downloads. The raw data was too broad (date wise) and specific (site wise) so needed wrangling to select the required years and map the sites (mainly cities) to their respective regions. There were a small number of missing values in the weather data which affected aggregating to obtain averages.

Regarding the scraping there were challenges when scraping Wikipedia articles specifically as they are inconsistently presented and do not use many “class” selectors. The resulting data-frame was also not exhaustive so some sites in the data were not assigned a region.

The scraping was a manually intensive process, but some automation was still possible. Creating a list that associates the data sites with their respective region was too difficult to automate so was achieved manually. There were 6 sites that didn’t correspond to the final data-frame. These were all manually assigned. Interestingly 2 of them were due to spelling differences which is a quirk of using the Māori language for place names.

The final data-frame consisting of the historical data for earthquakes, internal migration, and weather included quite a few missing values. These prevented effective plotting however, after cleaning it, the number of regions represented reduced to 13/16 with Nelson only having a row

for 2015 whilst Taranaki missed a row for 2014. Additionally, the measurements are averages that oversimplify the true nature of nature.

Techniques used

The csv files were renamed to reflect the context of their data. To aid in mapping the regions, some scraping of Wikipedia was performed to obtain a list of regions and their respective cities. Specific techniques include reading & writing csvs, visualising missingness, and working with data-frames (sub-setting, glimpsing, aggregating, merging). The missing values were 0.6% of the whole dataset so were omitted to prevent effects on calculating averages.

Data ethics

The dataset was created for anyone interested in a broad look at environmental factors affecting internal migration in New Zealand. Our team of four created the dataset for a group project at the University of Canterbury.

The data relates to natural events and internal migration counts so does not impact any individual with respect to privacy or discrimination.

The data was cleaned when producing the final dataset by omitting NA values which were present in the raw data.

The data will not be updated so acts as a snapshot for interested parties to get a general idea from.

What we achieved

Used geographical regional boundaries to aggregate a count of earthquakes per year within our specified year range. Distilled a complex geographical file wrangling process into an easy-to-interpret output. Wrangled weather data from StatsNZ into a simplified format that then joined with the earthquake and migration datasets. Managed to incorporate Sea Level Rise data into the final dataset, in a way that is customisable by the user to investigate different searise scenarios. This allows for an investigation into sea level rise as a driver for migration. Wrangled the data exported from statNZ to a connectable format with other environmental factors by aggregating districts into regions as the final output. Joined and plotted the final dataset using Julia for brief exploratory analysis.

What we failed to achieve

While there was a process of learning required to wrangle the earthquake data, the final output is an aggregated count of quakes per region. There are other dimensions of an earthquake that were lost in this output such as earthquake depth (which contributes to how violent the shaking feels). Further development of this research endeavour could expand on the earthquake data with possibilities such as scraping the web for sentiment about earthquakes in each region. I

would also recommend to future researchers that they investigate other natural disasters such as flooding or slips, as possible reasons for people movement in and around the country.

The complexity of the weather data is somewhat lacking. In retrospect, extremes in weather are more important than averages when analysing drivers of internal migration. The wind gusts were extremes so matching the theme with min/max temperatures and min/max rainfal would have resulted in a more dynamic and useful final dataset. This would also mitigate the negation of seasonal changes that occurred from averaging although ideally seasonal breakdowns could have been included.

Attempted to use a Julia API with StatsNZ API to obtain data on economic changes in a region. While this data is not an exploration of physical landscape changes to people's environment, it would have benefit to researchers, so that they could account for the effects of economic drivers of migration.

Could not change the x axis data type into string to create a better graph and could not include more valuable data into the migration dataset as expected initially.