Agricultural Activity and Water-Quality
Project Report

Authors:
Henri Liswoyo - 83967246 (hli173)
Alzen Punio - 11224000 (asp85)
Moses Velano - 83396373 (mbv16)
HanByeol Yang - 14868742 (hya62)

Group:
“Status 418: I’m a teapot”

DATA201 Group Project Report

Intro

Agricultural activity is linked to increasing contaminant concentrations in rivers and groundwater. For example, nitrate leaching and runoff. This occurs via nitrate fertiliser application, and nitrate in livestock urine (Source: Managing farm nitrogen | Waikato Regional Council). Drinking high nitrate water can cause negative human health effects, such as infant death via blue baby syndrome (Source: Blue babies and nitrate-contaminated well water. - PMC (nih.gov)).

Our group project aimed to collate, wrangle, and explore links between agricultural activity and water quality data to find a relationship between them. Moreover, create visualisations and a format to present the results.

Data sources

Stats NZ was used to source the following information:

- Horticultural activity data, broken down by area extent dedicated to each subcategory of horticulture within each region, by each year. E.g. The land area dedicated to apple farming.
- Livestock populations, broken down by species within a region, by year. Especially important to us were the population of beef cattle and dairy cattle.

Ministry for the Environment was used to source the following information:

- River quality data file for both E. coli and nitrogen indicators that includes the measurements of all the samples collected by regional councils taken from all of the monitored river sites across New Zealand.
- Land use area extent broken down by subcategory of agricultural activity in a region, per year. E.g. The land area dedicated to beef cattle.

LAWA was used to source the following information:

- Groundwater quality data file that includes the measurements of all the samples taken from all of the monitored well sites across New Zealand.

Reasons for sources

Stats NZ is the official data agency of New Zealand. The agency collects data from across different government agencies, and collates the data into databases that are easily accessible to the public. We selected this source, as Stats NZ is a regularly updated, and reliable source that carries data across many different topics relating to NZ. The data is available in a CSV format, which we are very familiar with. Stats NZ also advertised an API. We intended on using the Stats NZ API to scrape data, making it easier to collect large amounts of data. But, we later found that the relevant data we required was not available through their API. Their API is relatively new, and only contains data on COVID-19. Nonetheless, the API was one of the reasons why we selected Stats NZ as a source.

The Ministry for the Environment (MfE) is the NZ government's primary adviser on matters regarding the environment. Since our project relates heavily to land use and environmental

effects, the Ministry for the Environment would be a reliable source of information. The Ministry for the Environment collects data regarding land use and environmental effects in the nation to inform decision making, and policy changes. This data is in a publicly available database. We selected MfE as a source, since it is reliable, relates to land use in NZ, and contains data for the entire country broken down by year and by region. The MfE also makes their data available via an API. Using an API allows easier collection of large amounts of data, allows reproducibility of data collection within our notebook, and allows us to demonstrate our web scraping skills for this assessment.

Similar to the Ministry for the Environment, LAWA is one of New Zealand's sources of up-to-date environmental data. However, the groundwater quality data files in the Ministry for the Environment data service are limited and insufficient to meet the goal of our project; therefore, we used the available data from LAWA as it is complete, meaning it contains all of the necessary variables that we need.

Target use

The aim was for this data to be used as informative material in a digestible form. Our data is intended for use by people or parties that are concerned with the link between agricultural land use changes and changing water quality. Particularly, government agencies, concerned members of the public, and companies carrying out environment-impacting activities. These concerned parties may have a limited ability to carry out their own data manipulation, so digestible visualisation is important to making our data understandable.

This data will be used by the concerned parties to have a holistic view of the link between agricultural activity and changing water quality within their region. Alternatively, the data on water quality could be used separately to gain a holistic view of changing water qualities within their region. Perhaps, the user could link the trends to different water-quality-changing actions. The same can be done with agricultural activity, to identify trends in agricultural activity over time.

The data is intended to inform decision-making on water-quality impacting activities. Particularly, agricultural activity. On an individual scale, this could be used by individual members of the public to hold water-quality impacting activities to account. Or, on a wider scale, it could be used to inform decision-making for companies and/or regional councils.

Difficulties to overcome

A major difficulty of the project was finding a meaningful way to present our data so that it could be understood and analysed by the targeted users. Our initial plan was to produce a single data frame containing data for water quality measurements (e.g: nitrate concentrations), and census data on agricultural activity (e.g: livestock populations). These measurements were to be separated by region and by year. In the data frame, the combination of a region and a year was

considered a unique observation, and each unique observation had its own row. The measurements from that observation were to be presented in separate columns. E.g: If a user wanted to see the water quality and agricultural activity for Auckland in 2006, the user would scroll to the row for Auckland 2006 and look across the columns to see the relevant measurements. Roughly, it would appear as the following:

Key	**Region**	**Year**	**Groundwater nitrate concentration (g/m^3)**	**Dairy cattle population**
…	…	…	…	…
Auckland 2006	Auckland	2006	58.85	122234
…	…	…	…	…

Another difficulty we had was finding and combining different datasets that, to some extent, are recorded in a similar structure. We had trouble at this stage as most of the available datasets (primarily for water quality) did not contain the necessary variable (e.g. Region), exclusively containing records that were not observed and recorded yearly or only available up to specific periods. This impediment incurred temporal issues, prevented us from acquiring more relevant data, and additional time for finding packages to address a particular situation, such as latitude and longitude conversion.

Furthermore, there are different units for the results of river quality samples. This is due to different test methods conducted by regional councils and organisations. Therefore, to standardise the "units" column, we need to spend some time researching and understanding the given units, which results in additional time added to the wrangling phase of the project.

Moreover, a challenge that we have encountered as a team was the limitation of time throughout the years of observations. There were differences in temporal resolution where gaps between years of observations were visible. This resulted in having different intervals in data collection and some years considered as not assessed (NA). This challenge was handled by plotting each variable independently of each other and was visualised through a general trend overtime (time-series graph) using a library package called highcharter.

Since everyone in our group did different parts of our project, it was important to control our project version using GitHub. Since a few of us were not familiar with GitHub, controlling our project using GitHub was quite challenging. We spent around two weeks getting familiar with GitHub and made a simple manual about how to use git and GitHub through a terminal. Fortunately, We had few conflicts when we merged our branches, however, it was still challenging when I needed to resolve these troubles.

Our code needed to be reproducible, without the user making any effort to download additional files. This meant that the source data needed to be reproduced within the code. Initially, we attempted to use read_csv() on the download links for the source data. This worked for only a limited time, since the URLs would eventually expire. Therefore, we had difficulty making the code reproducible. We overcame this difficulty by hosting the files on github, making it accessible via a permalink

Techniques used

For agricultural activity, we began by scraping data available through API. The content returned by the API was in XML format, so xml packages in R were used to turn this into dataframe. Any data that could not be scraped was downloaded as a CSV. The download links for the CSV files were unique to each webpage session, and would eventually time-out. This made the link incompatible with the read_csv() function. So, for reproducibility of our ipynb, we hosted the files on github to use a permalink.

To join the tables, we defined each unique observation as the combination of the year and the region in which the observation was made. So the year and region columns were combined to create a “key” column to join the tables. The keys had to match to make them compatible for joining, so the gsub() function was used to standardise the naming convention of regions. E.g: Manawatū-Whanganui could be written as “Manawatu-Wanganui Region" which R would treat as a different region. This was standardised as Manawatu-Whanganui. The select and filter functions were used to remove irrelevant data. Such as: aggregate data relating to the entirety of NZ, and the count of farms within each region.

The dataframes were now cleaned and related by our new “key” column. We joined the data frames using the key column and then spread the data frame into a wide format. We produced visualisations of livestock populations from 2002 to 2019, across NZ.

For the water quality data file, initially, we took and converted the latitudes and longitudes from the river quality datasets we retrieved from the Ministry for the Environment via API. Then, we utilised the "reverse_geocode" from "tidygeocoder" package that uses different geocoding services such as OSM to locate and get the region where the river sites are located. We then write CSV files that contain geographical information.

We read and looked at each dataset in the Jupyter Notebook to get an overview of their structure and which columns contain NAs. Then we selected the necessary rows and columns, renamed them, and mutated their values. We chose the records from 2002 to 2019 as we decided to limit the period to have a sensible inference with the agricultural activity data. After that, we standardised the units with the indicators and some of the region names that were recorded unconventionally.

Using the cleaned data frames, we created four related data frames (tables) by combining particular variables as the key that uniquely identifies the coordinates of well or river sites.

These relational data frames help create plots such as the overall change in the average E. coli counts and nitrogen concentration. Furthermore, we were able to create visualisations that display the proportions of monitored sites under specific condition bands.

Achievements and failures

- Conversion of latitude and longitude (achieved)
    - Since specific datasets do not contain the "Region" variable, we used the "tidygeocoder" package that uses geocoding data to locate the area of the sites using, in our case, the latitude and longitude. The downside to this approach is longer processing time; nevertheless, we were able to get the region that allows us to connect the data frames we produced.
- Mapping (achieved)
    - We originally attempted to mean concentrations of contaminants in water using R. This proved to be unintuitive. We were initially unsuccessful with mapping using R. Henri was familiar with using ArcMap to produce choropleth maps, so ArcMap was used instead. Henri initially attempted to use the Intersect tool to produce statistics broken down by region. This was unsuccessful since the tool does not produce aggregate statistics, but takes the first result matching the region instead. The spatial join tool was used instead. The tool did not work as intended initially. It was found that the column containing values for measurements was being treated as a character type by ArcMap. Changing this column to a float produced the intended result. This could be exported as a shapefile, and uploaded as an interactive map on our *ArcGIS story map* web-page. So, mapping was eventually successful
- Water quality: accessed groundwater information (achieved)
    - Data files about groundwater quality are difficult to find, as most of the data services we encountered only have files containing limited and outdated information. Moreover, accessing the databases offered by government agencies and science institutes is complicated, as some of the databases require authentication and querying skills to retrieve the files. Nevertheless, we got the groundwater quality dataset from the LAWA website, which is sufficient for the aim of our project.
- Accessed Farm sizes information (achieved)
    - Data on farm sizes broken down by agricultural activity was required to produce statistics of livestock population density, and statistics of area extent. This was difficult to find. We suspected that this information may have been suppressed to protect business interests, similarly to data on livestock populations. Later, we found the relevant data on the MfE database. The API was used to scrape this data. This proved to be difficult, as the documentation was unclear on the format for an API query, and the query provided with the data only gave partial results. We resolved this by editing different parts of the query to get complete results. I.e. Removing a section of the query URL that specified returning only the first 5 results. This returned a result in XML format.
- Flexdashboard (fail)

    - Initially, we planned to create a dashboard using shiny and flexdashboard to display the findings from our data frames. Although we were able to create interactive plots, maps, and inputs, due to complexity, unfamiliarity, and time constraints, we could not utilise its potential to create an informative and interactive dashboard for our target users.
- Human health effects (fail)
    - We had originally planned on including data on the potential human health effects of poor water quality. Nitrate consumption is linked to colorectal cancer, and “blue baby syndrome”. If agricultural activity is linked to increasing nitrate concentrations in water, then it may have been possible to see a link between agricultural activity and the frequency of these conditions. We dropped this idea early on in our project, due to the lack of available data, and due to our method for workload distribution. Since we split into two subgroups working on one topic each, having a third topic complicated workload distribution and distracted us from perfecting our other focuses. This idea could have been possible if there was more data available, and perhaps if we had a more efficient workload distribution. Or, more time to work on this project.