

Disease surveillance using transaction data in New Zealand

Lijun Wang¹, Diego Enriquez Aparicio², Demival Vasques Filho³*

1 Te Pūnaha Matatini, Department of Statistics, University of Auckland, Auckland, New Zealand

2 Te Pūnaha Matatini, Department of Computer Science, University of Auckland, Auckland, New Zealand

3 Te Pūnaha Matatini, Department of Physics, University of Auckland, Auckland, New Zealand

 These authors contributed equally to this work.

* d.vasques@auckland.ac.nz

Abstract

Monitoring the spread of disease during or in advance of an outbreak can allow authorities to minimize harm to society and reduce the financial burden of such events on government. While infectious diseases still are a major threat worldwide, traditional methods of disease surveillance remain expensive and slow. In this context, due to the increasing ease of data collection accross multiple sources, big data analysis offers a complementary method of monitoring with great potential. Here, we investigate disease outbreak in New Zealand, using surface-level de-identified transaction data obtained from the Westpac New Zealand bank. First we look at the gastroenteritis case in the Hastings District in 2016 and, second, at the 2013 and 2014 flu season in the Queenstown area, a major tourist destination in the country. We show that it is possible to track the effects of the outbreak based on changes in the spending patterns in the affected area. In Havelock North, the disease caused the shutdown of business in the town, forcing residents to commute further for shopping, but spending less per transaction than usual. In Queenstown the beginning of the Winter season (not the peak of the flu itself), the Winter Festival and school holidays are the main drivers of how customers spend.

Executive summary

In this project we check the viability to predict, observe and track the spread of diseases using transaction data. Because the available dataset does not contain highly detailed information, such as, for instance, the exact product or service that is being purchased by the consumer, we refer to it as **surface-level transaction data**. On the other hand, the advantage of using such a set for disease surveillance is its scale, due to the large quantity of transactions made every day. We believe that it is possible to perform disease surveillance by finding changes in the spending patterns of the affected population.

First, we study the transactions that occurred in the Hastings District, from 2014 to 2016. Our intent is to observe variations in spending patterns caused by the outbreak of campylobacteriosis in the town of Havelock North in August 2016, due to contaminated drinking water. Our main findings are as follows:

- It would have been difficult to provide an early warning of an outbreak in this case. Unlike infectious diseases, where one must come into contact for spread, contamination in drinking water affects the population almost instantaneously.
- On the other hand, observing and tracking the effects of the outbreak after the initial infection event is found to be possible.
- In the second week after the contamination — the week after the boil water notice — spending returned to the previous levels everywhere in the Hastings district except for Havelock North. In Havelock North, where many people were affected, low levels of spending indicate a shutdown in local businesses due to owners and staff becoming ill.
- Havelock North residents were observed to travel to nearby areas to shop. The activity of Havelock North residents buying outside town increased, whereas the spending per transaction decreased. This could be explained by people commuting further to buy essentials only, spending less than average in each transaction.
- Furthermore, the proportion of spending at health-related merchants also increased in the second week for the entire district.
- The results also reveal that it is possible to identify travel (and changes in travel patterns) by Havelock North residents into the wider Hastings district. This type of information would be useful in developing models of disease spread.

Secondly, we analyze the transactions in the Queenstown area in 2013 and 2014, from April to November. The goal in this case was to observe the effects of the flu season. Although with only two years of records, it is possible to see the formation of an annual pattern for the spending behavior in that area, during the aforementioned period.

- Transactions and spending on health increase significantly in the beginning of June, coinciding with the start of Winter and the flu season.
- It seems that the search for medicines is more consistent in the beginning of the season than when the flu has its peak. Transactions and spending in health do not reach the same levels in the latter as for the former.
- Spending on health then decreased in the end of June and during the whole month of July, being replaced by consumptions in alcohol- and entertainment-related merchants. This coincides with the Winter Festival (June) and school holidays (July).

These results suggest that it is possible to use transaction data for disease surveillance, although use for early warning of disease outbreaks seems less likely. In addition, the data shows promise to be of use for:

1. Civil defence modelling and response (including pandemics), of interest to regional councils, ESR and the Ministry of Health;
2. Transport modelling and planning, of interest to Auckland Transport and NZTA, but also MPI for biosecurity;
3. Urban economic geography (investigating agglomeration effects), of interest to ATEED, Auckland Council, MBIE, and Treasury.
4. Social investment and well-being, especially the Social Investment Agency but also other social sector agencies, and Treasury.

1 Introduction

In the new era of big data and social media, new developments are improving our ability to predict disease outbreaks. Identifying big data, either in the form of clinical visits and pharmaceutical purchases or as search queries and social media, is becoming an important instrument for recognition, containment, and treatment of many diseases [1].

In a recent review, focusing on search queries and social media for disease surveillance [2], the majority (66%) of the 32 papers examined report that social media-based surveillance had comparable performance to existing surveillance programmes. Such methods are effective and allow for rapid detection. On the other hand, the potential for false positives and false negatives is a weakness, indicating the necessity to reduce surrounding noise. Most of the authors (75%) recommend the use of social media to support existing surveillance programmes.

More specifically, Ginsberg et al [3] find correlations between queries and physician visits and accurately estimates weekly influenza activity with a lag of one day. In [4], weekly Internet search query surveillance data reported by Google Flu Trends were found to accurately predict epidemic peaks 4-6 weeks in advance in metropolitan Melbourne (Australia). Apart from influenza, [5] compares eye disease diagnostics with social media data using Spearman rank correlation. The authors found that Internet-based search engine and social media post has a strong association with the occurrence of clinically diagnosed conjunctivitis as seen in electronic medical records (EMR). Furthermore, social media might be used earlier than the physician, and timing is crucial in detecting an epidemic [6].

Another approach is the use of spatial data for the surveillance of infectious diseases [7]. Technologies like social media, mobile phones, web search data, medical claims and pharmacy transactions produce spatial data which might help to understand the spatial distribution and dynamics of critical diseases. Emerging and Infectious Disease Outbreaks (EIDO) like Zika require to find new ways to predict and monitoring and presents new opportunities for methods. A good example is the Digital Participatory Surveillance where volunteers report their symptoms, signs and risk factors [8]. Other data sources that can be used to characterize an outbreak include online reports released by ministries of health, surveillance systems and the World Health Organization (WHO). Such documents have proved valuable to understand patterns for epidemic emergencies [9].

The use of financial data has also been used. This approach has drawn growing attention in the field of disease surveillance and many studies in the area have used such data in an attempt to better observe and predict the outbreak of diseases. However, like other methods, the use of financial data for disease surveillance has been applied in a range of different circumstances. Looking specifically at spending on influenza medicines, studies suggest that over-the-counter (OTC) sales may be a sensitive and early indicator of the spread of influenza [10, 11]. The first, an early work performed on 1979, claims that spendings on cold remedies increase an average of 185% above the baseline during peak influenza activity. The second shows a 90% correlation between flu-remedy sales and physician diagnoses of acute respiratory conditions with a lag of three days.

More recently, a comprehensive review of more than 3200 articles from different databases [12], mainly focused on severe respiratory and gastroenteritis infections, found that drug sales data and over-the-counter (OTC) drug sales analysis have the potential to forecast influenza-like illness (ILI) 1 to 3 weeks ahead. Almost all reviewed studies showed a high correlation between OTC and traditional surveillance data. The study also points out the requirement of electronic information systems for adequate monitoring of drug sales.

Systems that enable real-time disease surveillance, like the one presented in [13] includes spending records of individuals who are known to be infected with at least one

disease. The idea is to identify purchasing patterns that correlate to a population with a specific disease. They analyze data containing information like merchant name, merchant type, item or service purchased, account and customer information. From that, it is possible to compare the respective purchases with indications of infection associated with at least one disease.

However, dealing with big data sources like healthcare and hospital records, volunteer self-reports, Internet searches, social media, and mobile phones pose many challenges [15]. Here, we look at possibility of using surface-level transaction data for disease surveillance, to overcome these challenges. By surface-level transaction data we mean that this type of data do not contain detailed information either about the kind of products are being acquired — unlike OTC data — nor medical records associated with cards and accounts of people buying such products.

First we look at financial transaction in the Hastings District in New Zealand. In August 2016, there was an outbreak of campylobacteriosis in the town of Havelock North due to contaminated drinking water. Based on the report of the inquiry of this contamination [16], it is highly likely that heavy rain on the 5th and the 6th of August in this area caused the contamination. After tests and confirmation of the contamination, a boil water notice was released via conventional and social media on 12 August 2016. Over 1,000 cases were reported, but an estimated number of 5,500 residents, out of a total population of 14,000, had become ill with campylobacteriosis. Thus, our focus of study in this case is to track the differences in transaction patterns and spending behavior between the contamination time period, especially the preliminary stage of contamination before the announcement, as well as other time periods.

We also look at financial transactions at the Queenstown area. Queenstown, a town of about 15,000 residents, is a major center of adventure tourism in New Zealand, including Winter sports. It receives a large number of tourists also during the flu season, and our goal in this case is to analyze the impact of the flu season in the spending patterns in the area.

The rest of this work is as follows: in Section 2 we discuss in detail the available data, their limitations and the techniques we used to analyse them. Results and discussion of our analysis and findings are in Section 3. Finally, we present our conclusions in Section 4.

2 Data and methods

2.1 Dataset

Financial transaction data were mined from the data set provided by Westpac for two specific areas in New Zealand: the Hastings District, located in the Hawkes’s Bay region, and the Queenstown area, in the Otago region.

The datasets were built using four sources comprising EFTPOS and Scheme card transactions (credit and debit card transaction records) from Westpac’s core system joined to customer data via an account reference table. A privacy impact assessment was conducted during the scoping of data request. Different data masking treatments were applied to fields of the dataset to prevent identification of individuals or organizations. Information about customer, account, card numbers and merchant has been hashed, using SHA256. On the other hand, spatial data, provided using Meshblock and NZ Postcode references, were truncated due to the small number of individuals or organizations in small urban areas. All variables included in the datasets can be seen in Table 1.

Due to Westpac’s card payments processing, the following type of transactions are part of the datasets (see a summary in Table 2):

Table 1. Dataset attributes/variables. Data masking treatment applied for POST_CODE_T, ACCOUNTX, CARDPANX, MERCHANT_IDX, CUST_NUMBERX and POST_CODE_C

Attribute	Description
TRAN_CODE	Transaction code — type of transaction
TRAN_AMT	Transaction amount
FIID	Financial Institution Identifier (WBCO, PBK1, TSB1, KWB1, ASB1, NZCU, MCD1, VIS1, PAYM, EFT1, BNET, VER1, VISA, AEGN, DPS1, MQIB)
CARD_PRESENT_IND	Indicates whether the card was present at the moment of the transaction
MCC_LVL1	Merchant classification code - Level 1
MCC_LVL2	Merchant classification code - Level 2
POST_DATE	Transaction processing date
POST_CODE_T	Transaction postal code
TRAN_DATE	Transaction date
ACCOUNTX	Hashed account code
CARDPANX	Hashed card code
MERCHANT_IDX	Hashed merchant code
TRAN_TIME	Transaction time — 10 minutes increments
CARD_BIN	Bank identification number of credit/debit card
TERM_COUNTRY	Country (all values for New Zealand)
OCC_CODE	Occupation Code
AGE	Customer age (band)
CUST_TYPE	Customer type (individual or organization)
CUST_NUMBERX	Hashed customer code
INVOLMENT_ROLE	Primary or additional customer
INVOLMENT_TYPE	Private, joint, as trustee for or non-personal account
ACCOUNT_NUMBER_FORMAT	Credit card (V) or bank account (C)
MESHBLOCK_ID	Meshblock ID number
POST_CODE_C	Customer postal code
PROC_DATE	Dataset processing date
SOURCE_SYSTEM	Transaction log file (PTLF or TLFx)

1. All credit and debit card transactions, to a retail merchant, conducted using a Westpac issued card. This includes present and not present card transactions (like online purchases).
2. All credit card transactions (not debit), regardless of issuing bank, processed by Westpac, where Westpac is the merchant acquiring bank for the organisation receiving funds.
3. ATM channel transactions for Westpac, other bank and specialist card issuers conducted on Westpac ATMs.
4. ATM transactions where Westpac cards were used in non-Westpac ATMs.

Lastly, the following records were excluded:

Table 2. Card transaction content. Types of card transactions captured by Westpac that comprise the dataset of the project. It excludes non-Westpac debit card activity on Westpac acquirers and both credit and debit card activity on non-Westpac ones.

		Issuer	
		Westpac	Not Westpac
Acquirer	Westpac	Credit/Debit	Credit
	Not Westpac	Credit/Debit	None

- Customers and accounts owned by customers identified as aged under 18 years.
- Identification of international locations in which transactions take place.
- Transactional activity and information relating to wholesale banking arrangements.
- Transactional activity and information relating to entities which have opted out.

2.1.1 Limitations

As stated before, some information in the data set was either encrypted, as for account number, card number and merchant identification or truncated, as with some post codes, in order to preserve privacy. For the latter, just the first two digits of the post code were present in locations with a small number of customers.

Moreover, not all transactions made in the aforementioned regions are available for analysis in this project. An estimated 35% of the total number of transactions in New Zealand are captured by Westpac. Table 2 shows the situations in which card transactions are present in the data. When Westpac is the issuer of the card, both credit and debit card activity are always present in the data. Otherwise, credit only card activity is present when Westpac is the acquirer and no activity at all when both issuer and acquirer are not Westpac.

Finally, although ATM transactions (withdrawals) are included as part of the data set, it is impossible to track where subsequent expenditures in cash occur. This is particularly true, as we will see later, for the town of Havelock North, where ATM transactions are unusually frequent.

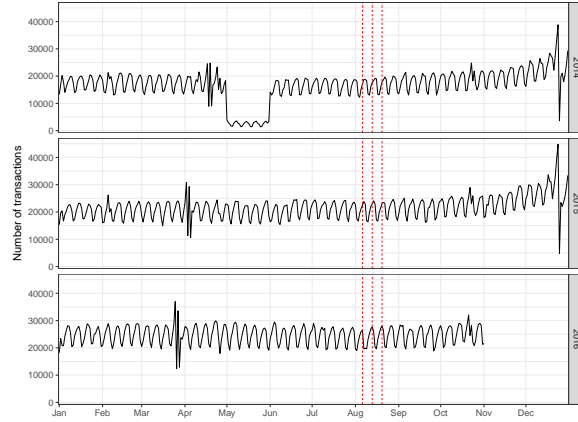
2.1.2 Havelock North

The dataset for the Hastings district has a little over 30 million transaction records from January 2013 to October 2016. About 35% of the dataset (near 10 million) contains Westpac customer information such as a hashed customer number, age, type, involvement and postal code.

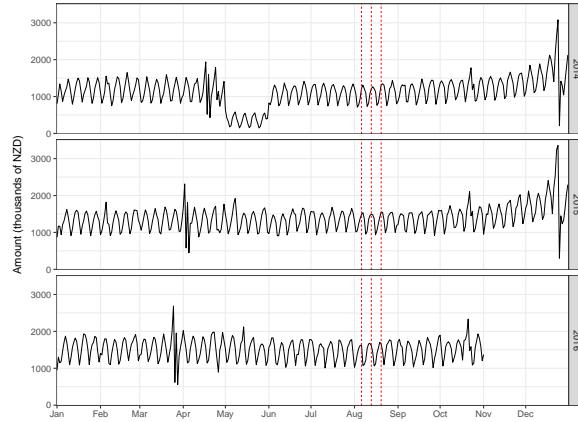
Due to the postcode generation process explained before, the transaction and customer postcodes show noise with some locations that do not belong to the area of interest. The Havelock North postcode 4130 only represents 1% of the transactions and 10% of the customers. Surrounding and bigger towns like Hastings (4122) and Napier (4110 and 4112) represent 11% and 22% of the transactions, and 10% and 34% of the customers respectively. But, 29% of records (almost 9 million transactions) have postcode value of 41, covering Hastings (which includes Havelock North), Napier and Wairoa areas. In summary, analysis performed in this dataset was for the following postcodes: 4130, 4122, 4110, 4112 and 41.

About 85% of the transactions are between 0 and 100 NZD and 13% between 100 and 500 NZD. Few transactions have negative values, transactions between 1000 and 10,000 NZD are rare (1%), and there are few extreme outliers greater than 10,000 NZD.

Figure 1 shows an overview of the number of transactions and the total spending across three years, from 2014 to 2016. There is a marked weekly trend with peaks on Fridays and lows on Sundays and Mondays. The trend seems steady all the year with an increase in transactions and spending towards the end of each year. Also, there is a slight trending increase across all years.



(a) Number of transactions



(b) Total spending

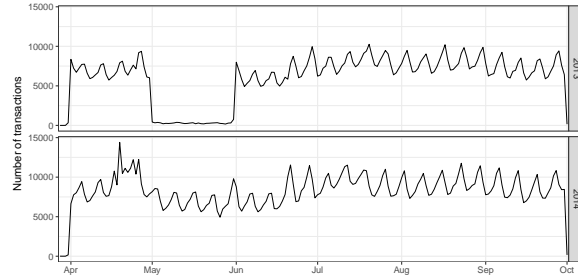
Fig 1. Transactions and spending in the Hastings District. (a) Number of transactions by year and (b) total spending amount in thousands of New Zealand Dollars (NZD). Total spending is the sum of all transaction values of the day. Both graphs show almost identical behavior with a marked weekly seasonality. The data corruption reported in May 2014 is evident. Outliers correspond to Easter, Labour Day and Christmas respectively. The dashed lines are a reference to the incident first two weeks, from August 5th to 19th.

The prominent peaks observed each year match Easter, Labour Day, and the Christmas holidays. Easter and Christmas exhibit more dramatic behaviour with a considerable increase in numbers of transactions and spending the day before the holiday period, followed by a drastic drop during the holiday period. The significant low

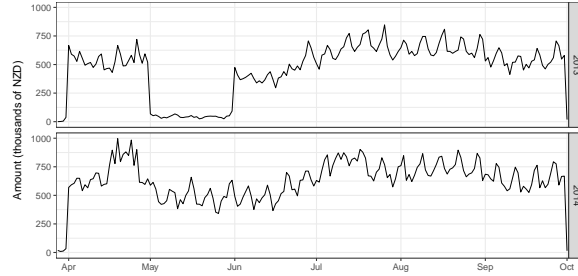
for the whole month of May 2014 for both number of transactions and spending is a result of data corruption in this month. Therefore, the records in this period are not used for any analysis throughout this work. Lastly, the three red dashed lines serve as reference for the dates when the August 20016 incident occurred, starting on August 5th with the heavy rain that caused the water contamination, August 12th, when the alarm was raised, and the the week after.

2.1.3 Queenstown

The Queenstown data is divided into two datasets with around 1.5 and 2 million transaction records for 2013 and 2014 respectively, from April to September. From that, about 14% and 21% of the dataset contains Westpac’s customer information. The same postcode generation process explained before applies in this case. For the 2013 dataset, Queenstown postcode 9300 represents 32% of the transactions and 2.1% of the customers. There are 37% of records (more than half million) with a postcode value of "93", covering Queenstown, Arrowtown and Cromwell areas. In the case of 2014, postcode 9300 represents 33% of the transactions and 3.4% of the customers. Postcode 93 comprises 38% of the dataset.



(a) Number of transactions



(b) Total spending

Fig 2. Transactions and spending in the Queenstown area. (a) Number of transactions by year and (b) total spending amount in thousands of NZD. Here again the weekly seasonality is noticeable. Differently from the Hastings dataset, in this case is possible to see the increase of transactions and spending during Winter.

The dataset shows a marked weekly variation with peaks on Fridays and Saturdays, going down on Tuesdays and Wednesdays (Fig 2). No outliers are present, but an increase in transactions and spending occurs during Winter. About 82% of the transactions are between 0 and 100 NZD and 16% between 100 and 500 NZD. Few transactions have negative values — most likely refunds —, transactions between 1000 and 10,000 NZD are rare (0.5%), and there are few extreme outliers greater than 10,000 NZD. As with the Hastings District dataset, there is a corruption data issue for the

Queenstown area, although instead of May 2014, it occurs in May 2013.

2.2 Methods

2.2.1 MCC grouping

The original data contains a large number of Merchant Category Code (MCC) — the business scope of the merchant — categories. There are 370 level 2 categories grouped into 71 level 1 categories in the original datasets. Moreover, many of them are either duplicate or have a redundant description. For both reasons, in order to simplify our analysis, new labels were assigned to the MCC categories, grouping them into 20 new categories based on their original descriptions, as shown in Table 3.

Table 3. MCC new labels.

Airlines	Food	Other transportation
Alcohol	Gas	Stores
Automotive	Government	Supermarket
Beauty	Health	Transportation
Education	Lodging	Utilities
Entertainment	Others	Wholesale distributors and manufactures
Financial	Other services	

As part of our proposed approach, we assume that the outbreak of a disease could impact some of these new labels in opposite ways. So we look at three of them specially: Health, Alcohol and Entertainment. We call the impact caused in the first as positive impact. That means the we expected an increase in spending in health products and services due disease spreading. On the other hand, the outbreak would cause a negative impact on the last two, meaning that we expected a decrease of spending in businesses related to alcohol and entertainment.

To magnify the difference in the spending patterns related do those three MCCs, we calculate the ratio of positive impact MCC (Health) with regards to the negative impact MCC, given by

$$R = \frac{M_{health}}{M_{alcohol} + M_{entertainment}}$$

where M is the total spending (or total number of transactions) for each MCC.

2.2.2 Moving averages

The traditional use of moving averages is that at each point in time we determine averages of observed values that surround that point. In other words, we are creating a series of averages of subsets of the entire dataset. We do so to smooth out the short term fluctuations in our data, due the aforementioned weekly patterns of people’s spendings. By minimizing such fluctuations, we can better identify general spending trends.

There are many types of moving averages, e.g simple, cumulative, weighted, exponential and so on. For our purposes, we use simple moving average, which is calculated according to

$$\overline{M}_t = \frac{1}{n} \sum_{i=0}^{n-1} M_{t-i},$$

where n is the size of the subset (number of observed values in the average) and M_t is the observed value in time t . For the entire work we use $t = 7$, for the days if the week, unless otherwise stated.

3 Results

3.1 Havelock North

Our first approach is to look at the broader Hastings District. We investigate the proportions of both number of transactions in Health, to check the change in activity in the period of the outbreak, and spending patterns. Fig 3(a) shows that in the two weeks after the water contamination (red dashes lines), the proportional activity within the Health label actually decreases when comparing to the previous weeks. Similarly, this behavior is seen for the spending proportion on Health (Fig 3(b)). One would expect otherwise in an event like the outbreak of a disease. However, transactions within the Supermarket label might be masking the results. Consumers could be buying a large amount of medicines in supermarkets. As we do not have detailed information of what is being purchased (like OTC data) and as our goal is to perform disease surveillance using surface-level data, supermarkets act as a 'black box' for us.

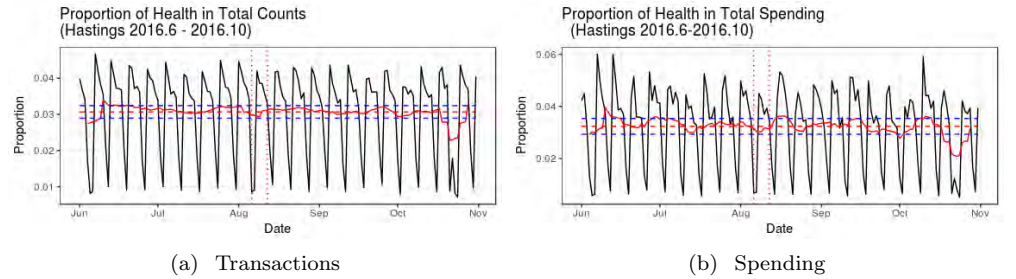


Fig 3. Proportion of Health related (a) transactions and (b) spending in the Hastings District. The weekly pattern is noticeable and we use moving averages as a trend indicator (red line). Blue dashed lines are the standard variation from the mean (red dashed line) of the moving average values. The second week after the contamination show a decrease in transactions but significant increase in spending. That indicates an increase in the mean amount spent per purchase.

The alternative is a more detailed analysis on industries that are more likely to suffer a stronger impact with the spread of the disease. Following this reasoning we investigate the assumption that Health merchants are part of the group that shows positive impact while Alcohol and Entertainment merchants are part of the group that experience a negative impact with the outbreak. The evolution of the ratio between positive and negative impacts, for transactions and spending, are show in Fig 4. Again, the ratio of transactions in the Health category decreases when comparing to the previous weeks. However the ratio of spending surges. That means that even though fewer visits were made to health-related shops, such visits turn out to have a relatively larger consumer spending than for alcohol- and entertainment-related merchants.

Due to the subtle differences in spending behavior for the entire district, we investigate further the transactions made only in the town of Havelock North, where the vast majority of the people affected by the outbreak lived. We look at the trends for each day of the week in order to check the behavior of number of transactions and spending separately. Fig 5 shows a range of two months before and after August 2016 with each day of the week in a different color. It also shows the moving average for each day with a window of eight weeks.

From Fig 5(a) it can be seen that Fridays and Saturdays are usually the days with the highest number of transactions. But the August 6th transactions show a shallow value for a Saturday. In fact there was a 'weather bomb' across Hawke's Bay area on this day where heavy snow and flooding provoked road closures, power outages among

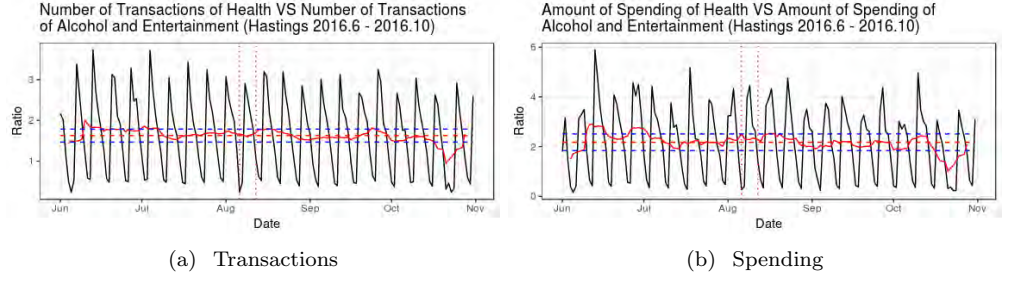


Fig 4. Ratio of Health over Alcohol and Entertainment in (a) transactions and (b) spending in the Hastings District. Moving average values are consistently above its mean in the weeks after the water contamination in a similar pattern as June and July, beginning of Winter season.

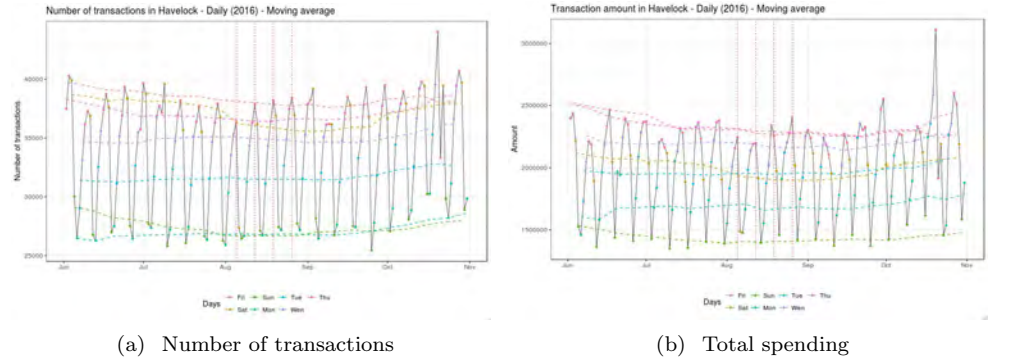


Fig 5. Average number of transaction and spending in 2016 for the town of Havelock North only. The moving averages are an indicator of the trend of transactions and spending for each day of the week. We can see the dramatic drop in transactions and spending on Saturday, 6th of August. This was the day after the contamination and when a 'weather bomb' hit the area.

other problems [17]. Most of the moving averages show a slight curve going down and reaching the lowest value around August and September to start to increase towards the end of the year. This behavior seems proper for the winter season in that area. Saturdays show a deep curve mostly because of the effect of the influential value on August 6th. Monday and Tuesday present unaffected behavior. An important fact to notice here is a group of low values for the number of transactions for the second week of marked in red dashed lines. Transactions on every day of such week is below their respective moving average showing a unexpected decrease in the level of activity.

Figure 5(b) shows a similar pattern for the total spending. The effect of the 'weather bomb' is visible, but the low values for the second week are not as significant this time. Although the number of transactions decrease, the spending does not decrease as much, indicating an increase in the mean spend per transaction. In other words, this may indicate that people are not willing to go out for shopping as often as usual, stocking up on more goods per trip.

We divided the dataset into four groups considering the location of the transactions and the location of the customers: non-local — non residents in Havelock North — customers and local customers making transactions in and out Havelock North. Figure 6 shows the behavior in each group using the number of transactions and total spendings. The dash lines represent the three weeks starting on August 5th. In figure

6(a) an apparent decrease in transactions can be observed for all groups, this is expected concerning the 'weather bomb' announced on August 6th. However, the following week, starting on Friday 12th, shows a more profound decrease only in the transactions made in Havelock North. This fact coincides with the report indicating a severe number of cases reported between 13th and 18th. Many victims of the illness had to take time off work or school affecting local businesses.

Unlike the number of transactions in Havelock North that decreased in the second week, transactions in the surrounding areas were not affected. This could mean that many businesses in town were closed that week as a result of people being sick (business owners or staff). On the other hand, transactions from locals buying outside Havelock recovered, growing to even higher levels than before the water contamination event. This suggests that locals had to shop in the nearby areas due to the avoidance or shutdown of local businesses. The spending of locals outside Havelock North, however, decreased in the second week (Fig 6(b)). Hence, the mean spending per transaction also decreased, indicating that in this period when local business were closed, people went outside to shop but spending less. For the other cases, the total spending in figure 6(b) reveal the same behavior as for the number of transactions, with the second week showing a small depression. The spending of locals in Havelock North being the most strongly affected. A qualitative summary of the correlation between transactions and spending and the variation on value levels for the all cases can be seen in Table 4.

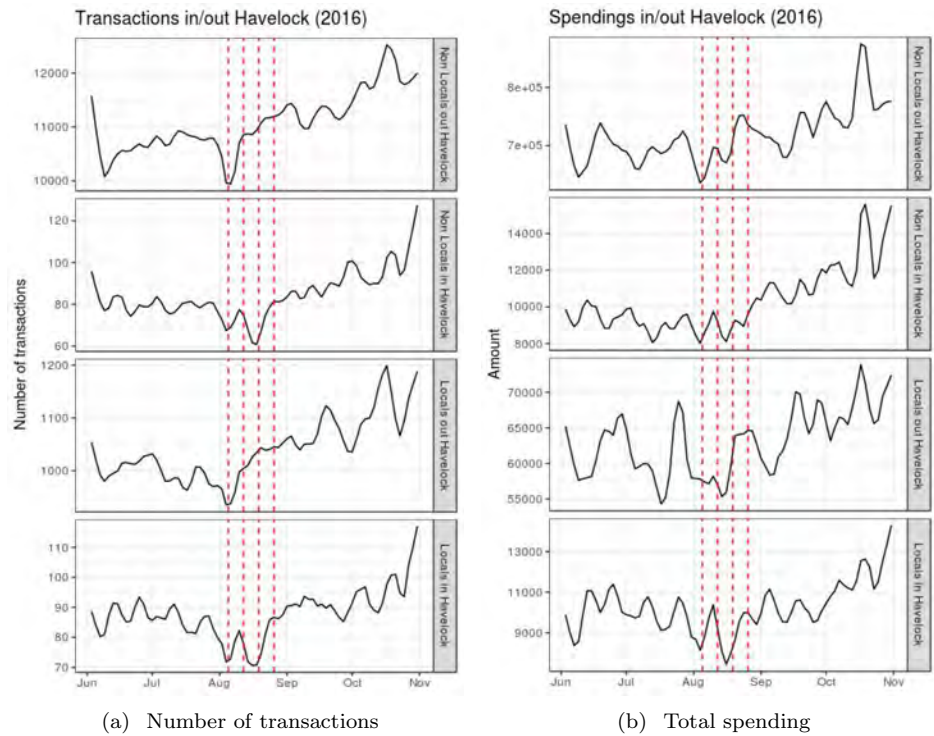


Fig 6. Comparison of (a) the number of transactions and (b) spending for local and non-locals in and out Havelock North. The drop in the second weekend after the contamination for both transactions and spending in Havelock North indicates the shutdown of businesses due to the outbreak. Locals had to buy more at the nearby areas increasing activity, however spending much less than usual per transaction.

Although the reference weeks show some effects that we can relate to the 'weather bomb' and the disease outbreak, it is possible to spot other periods with similar or more

Table 4. Correlations between transactions and spending drop, for local and non-locals of Havelock North spending in and out of town, in the first and second weeks after water contamination.

Resident	Shopping locale	Correlation	Variation
non-local	outside	positive	low
non-local	inside	positive	high
local	outside	negative	low
local	inside	positive	high

significant changes. These are likely to be related to particular events, and should be investigated on a case-by-case basis.

A close up of the transactions separated by MCC label reveals that almost 60% of these correspond to cash disbursements in the Financial label as shown in Fig 7. It is followed by 17% and 10% in Food and Alcohol respectively. The reference weeks of the outbreak (weeks 32 and 33), present subtle decrease in the number of transaction for all MCCs, without any in particular being responsible for the drop in the transaction activity. It is important to notice, however, that week 33 has one of the lowest levels of cash withdrawing in the entire year.

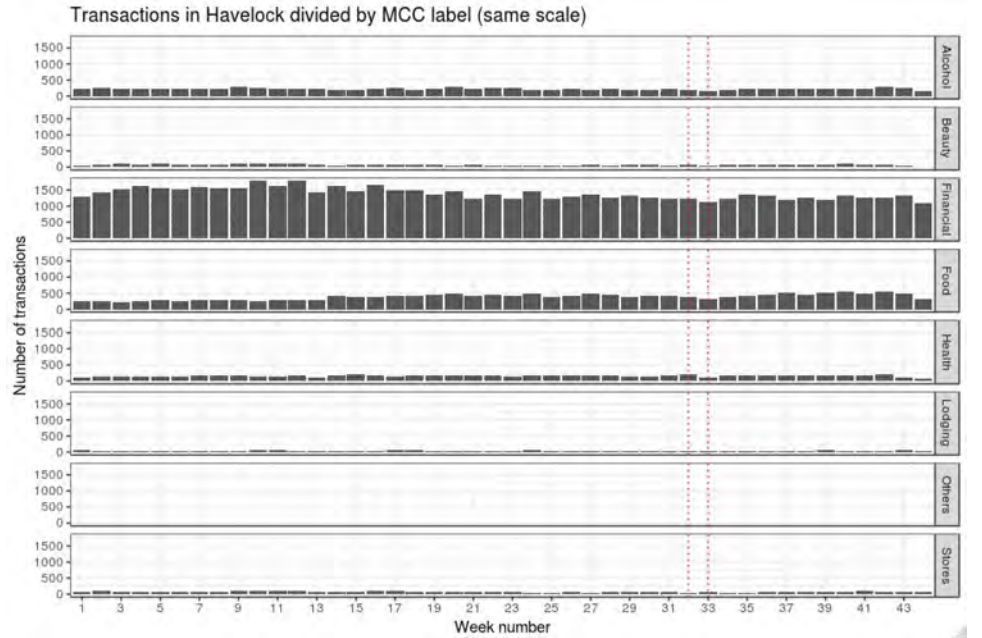


Fig 7. Transactions in Havelock North by MCC Label. The dominance of financial transactions (cash withdraw) is clear. In general all MCCs show a decrease in the number of transactions in the week after the contamination.

3.2 Customer groups in Havelock North

In this section we dig deeper to get more detailed information about the spending behavior of people from the town of Havelock North. The dataset contains information about whether a transaction is associated with a private or a joint account, primary or additional customer, and so on, but it has no information about which customers are

related to other customers. It is possible to make the assumptions that transactions and spendings would depend on the customer's relationships as it is common for families to share cards and bank accounts. Therefore operations with one card can be related to necessities of more than one customers. Also, some accounts can be used for specific purposes, like savings or emergencies, while others for day to day consumption.

When a card or an account is associated with more than one customer (excluding organizational type customers), it can be assumed that these customers have some relationship. Applying this reasoning, 14,127 customers were assigned to 7033 groups (from week 1 to 43 of the year 2016) that can now be used to explore and visualize the trend in transactions and spending of related customers.

Assuming that customer postcode 4130 means that the customer is a Havelock North resident (or a local), 5% are pure locals, and 1% has at least one local member while 94% are not from Havelock North. Focussing on the pure local groups, that is about 350 groups, we obtained the number of transactions per week and plotted the distribution of the groups depending on the number of weeks with transactions. Fig 8 shows that 140 groups have transactions during all the weeks of the year (43 weeks of data for 2016), 117 have between 1 and 4 weeks with no transactions (39 to 42 weeks), and 93 groups have more than 4 weeks with no transactions.

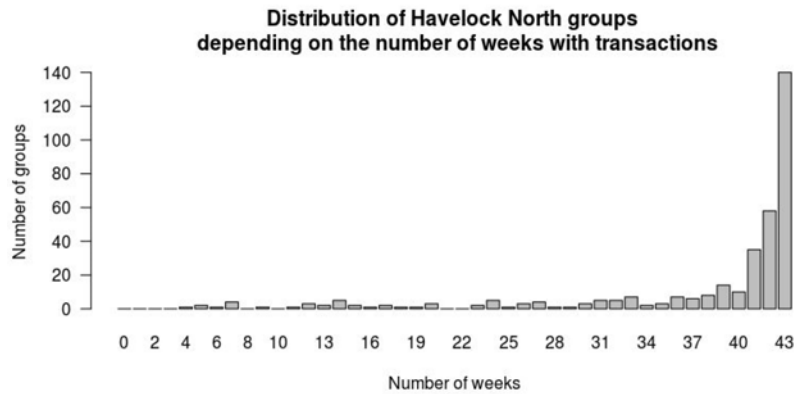


Fig 8. Distribution of transactions per week. The vast majority of identified groups in the Hastings district have transactions in almost every week.

Going further with the analysis, we consider frequent groups the ones that have more than 38 weeks of transactions, a total of 257 groups. The study shows that being these groups frequent customers and residents of Havelock North, only 4.5% of the transactions are made in Havelock North, but 40% have postcode 41, which may contain Havelock North's transactions as well. We see the same pattern found in the previous analysis of Havelock North residents doing most of their transactions in other towns.

If we only consider the transactions done in Havelock North by pure Havelock North groups we observe that the majority have few weeks with transactions during the year. Fig 9 shows a big cluster of groups with transactions in 15 or less weeks. As we saw in the last section, a big portion of the transactions in Havelock North correspond to ATM cash withdraws. If we remove such transactions, the effect on the distribution is notorious. Fig 10 shows that more that 150 groups (48%) do not perform any other operations during the year. In other words, the majority of the residents of the town still prefer cash over cards for their purchases.

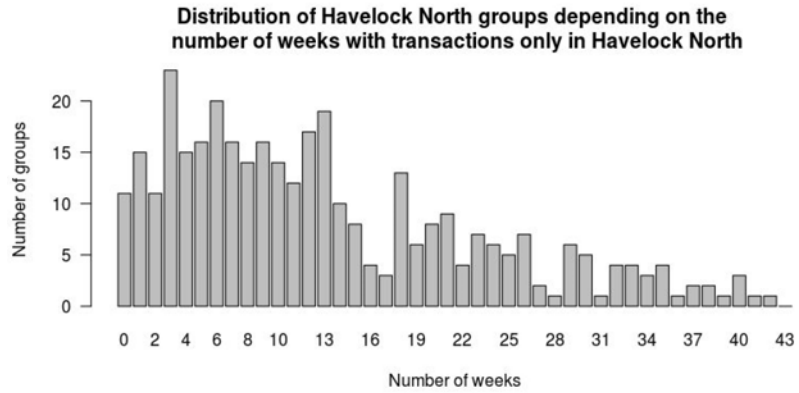


Fig 9. Distribution of transactions per week of Havelock North groups. The use of cards by the groups of Havelock North is much smaller than the use by groups in the entire district.

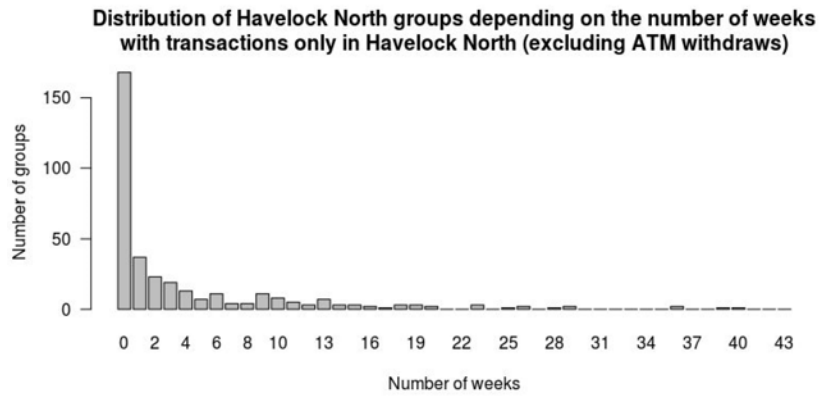


Fig 10. Distribution of transactions per week in Havelock North by local groups excluding ATM withdraws. Cash withdrawing is the major type of transaction in Havelock North. When this data is excluded, the activity in card transaction is very low.

3.3 Queenstown

The analysis of the dataset for the Queenstown area is similar to the analysis of the Hastings District. However, it is important to notice that for the former, there is not a short period of time for the occurrence of the disease like for the latter. The flu season is expected to be happen between the months of May to September. Moreover, the peak of the flu was in August (in 2013) and September (2014) [18, 19]. Due to this reason we focus on the proportions of transactions in the Health category and its ratio with regards to Alcohol and Entertainment categories.

Although it is hard to infer a long term seasonal trend in a data set containing only two years, some observations and assumptions are still viable. Number of transactions and value of spending start to grow in June, the beginning of winter, reaching their maximum values in July and August (Fig 2). This is an expected behavior since winter sports are a big attraction in Queenstown.

Interestingly, in July and even August (peak of flu for 2013) through September (peak for 2014) have levels of proportions of transactions and spending lower than in June, when they reach their top. The red line in Fig 11 shows the trend (moving

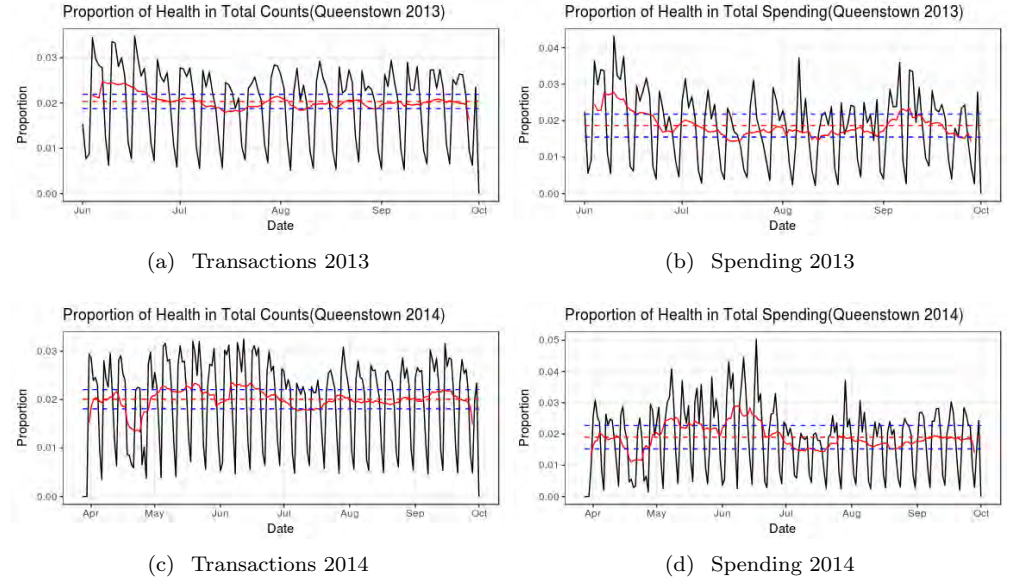


Fig 11. Proportion in Health in 2013 and 2014 for the Queenstown area. Proportion of (a) number of transactions and in (b) spending show that, with even only two years records, activity and expenditures grows in the beginning of June, when the Winter season starts, and drops in July, specially during school break.

averages) for both transactions (Figs 11(a) and 11(c)) and spending (Figs 11(b) and 11(d)) proportions in Health merchants. Horizontal dashed blue lines are the interval from the moving averages mean value to its standard deviation. The trend clearly shows the increase in the values in June, always reaching levels above the standard deviation. Although early Winter season is not the peak of the flu season, the first case might be leading to a rush for medicines shopping.

These results are supported by the ratio of transaction and spending in Health with regards to Alcohol and Entertainment (Fig 12. This ratio peaks at the beginning of June and fall to a minimum at the end of June and entire month of July, coinciding with the Winter Festival and school holidays respectively.

It is important to note that, according to the influenza reports [18,19], the flu season in both years were relatively mild when compared with previous years. Nevertheless, it seems that the spreading of flu, as for whole Winter season, does not exert a major influence on the economy of the tourism-focused town of Queenstown. Spending during Winter increases in the area and the peaks of flu happened outside the period when the town is usually busier, from the end of June to the end of July.

4 Conclusion

In this work we check the viability of disease surveillance by using surface-level transaction data. This type of data does not include detailed information as over-the-counter data and health reports. On the other hand, its scale (large number of transactions every day) gives an promising advantage for the study of patterns and behaviors of a population.

Our data analysis using transaction in the Hasting District show the possibility of tracking changes in spending patterns due to the outbreak of gastroenteritis in the town of Havelock North in 2106. It was possible to see the decrease of transaction activity

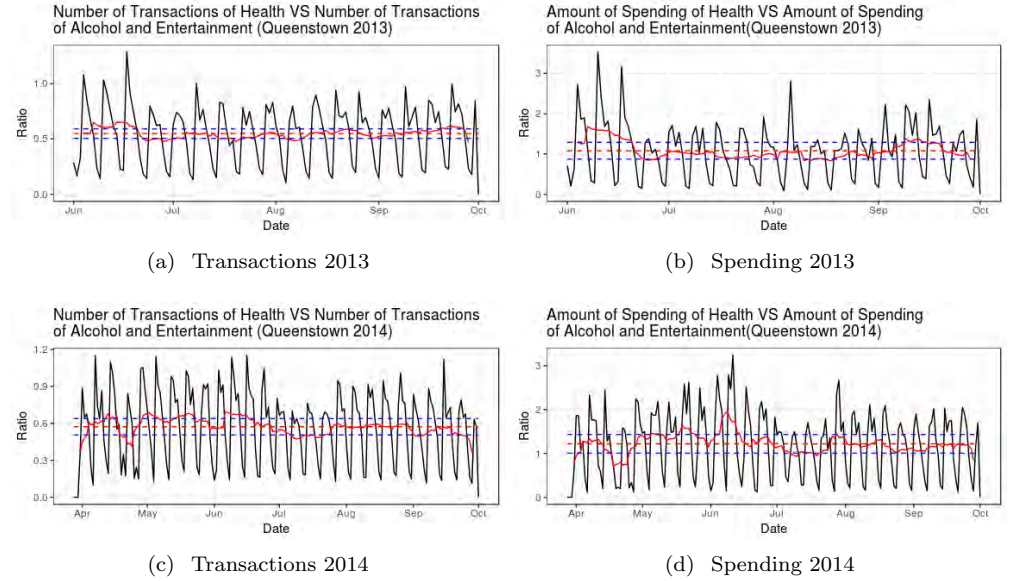


Fig 12. Ratio of Health (a) transactions and (b) spending with regards to Alcohol and Entertainment in 2013 and 2014. The ratios are much higher in June, coinciding with the beginning of the flu season and reach lowest levels in July, during school holidays.

and spending in the week following the water contamination, when a large number of residents became ill. The data indicate businesses shutdown in the period, affecting the local economy and the spending pattern of the residents. Increase in the number of transaction of locals, made outside Havelock North, show that people had to commute further for shopping. At the same time the amount spent decreases significantly. In other words, residents were probably just buying essentials, as the relative amount spent per transaction reached low levels.

The proportion of health-related transactions and spending, and the ratio of Health with respect to Alcohol and Entertainment labels, also for transaction and spending point out to the same conclusion as mentioned above. Although the number of transactions had not changed significantly in the period of the outbreak, spending had a meaningful increase in the week after the contamination. Hence, consumers were buying more, on average, per visit to health-related merchants.

Lastly, for the Hastings District dataset, it is important to notice the curious fact of the effect of the weather in the spending behavior. The water contamination was said to be due to the heavy rain that hit the area on the 5th and the 6th of August. On the 6th, a Saturday, the number of transaction and spending went down to levels compared to Mondays, the day of the week with historically the least activity in Havelock North.

For the Queenstown area dataset we looked at the effects of the flu in the spending patterns. Even with only two years of records, it is possible to detect an annual pattern. More years of data would be necessary to confirm the transactions and spending in health-related merchants reach maximum levels in the beginning of June. A possible reason for that is the start of Winter and of the flu season. Even though flu is peaked in August (in 2013) and September (in 2014), consumers may start buying medicine when the season starts.

Another clear pattern that could be confirmed with more data is the drop in the proportion of transactions and spending in Health starting in the end of July and reaching minimum levels in July. The behavior of the ratio of health over alcohol and

entertainment activity and expenditures are similar. Both results support the idea that the Winter Festival (end of June) and school holidays (in July) are the main factors that contribute to the spending patterns in the area during the Winter period.

5 Outlook

It is unlikely that a surface-level data set would have allowed identification of the Havelock North outbreak prior to its discovery by public health officials, partly due to the confounding reduction in transactions associated with the storm. However, the data set revealed significant shifts in spatial transaction patterns after August 12, which demonstrates that it will have considerable utility for other government agencies. In particular, the data will have use for:

1. Civil defence modelling and response (including pandemics), of interest to regional councils, ESR and the Ministry of Health;
2. Transport modelling and planning, of interest to Auckland Transport and NZTA, but also MPI for biosecurity;
3. Urban economic geography (investigating agglomeration effects), of interest to ATEED, Auckland Council, MBIE, and Treasury.
4. Social investment and well-being, especially the Social Investment Agency but also other social sector agencies, and Treasury.

References

1. A. Sommer, "The Utility of "Big Data" and Social Media for Anticipating, Preventing, and Treating Disease", *JAMA Ophthalmology*, vol. 134, no. 9, p. 1030, 2016.
2. T. Bernardo, A. Rajic, I. Young, K. Robiadek, M. Pham and J. Funk, "Scoping Review on Search Queries and Social Media for Disease Surveillance: A Chronology of Innovation", *Journal of Medical Internet Research*, vol. 15, no. 7, p. e147, 2013.
3. J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski and L. Brilliant, "Detecting influenza epidemics using search engine query data", *Nature*, vol. 457, no. 7232, pp. 1012-1014, 2008.
4. R. Moss, A. Zarebski, P. Dawson and J. McCaw, "Forecasting influenza outbreak dynamics in Melbourne from Internet search query surveillance data", *Influenza and Other Respiratory Viruses*, vol. 10, no. 4, pp. 314-323, 2016.
5. M. Deiner, T. Lietman, S. McLeod, J. Chodosh and T. Porco, "Surveillance Tools Emerging From Search Engines and Social Media Data for Determining Eye Disease Patterns", *JAMA Ophthalmology*, vol. 134, no. 9, p. 1024, 2016.
6. M. Deiner, T. Lietman and T. Porco, "Uncertainties in Big Data When Using Internet Surveillance Tools and Social Media for Determining Patterns in Disease Incidence—Reply", *JAMA Ophthalmology*, vol. 135, no. 4, p. 402, 2017.
7. E. Lee, J. Asher, S. Goldlust, J. Kraemer, A. Lawson and S. Bansal, "Mind the Scales: Harnessing Spatial Big Data for Infectious Disease Surveillance and Inference", *Journal of Infectious Diseases*, vol. 214, no. 4, pp. S409-S413, 2016.

-
8. C. Pagliari and S. Vijaykumar, "Digital Participatory Surveillance and the Zika Crisis: Opportunities and Caveats", *PLOS Neglected Tropical Diseases*, vol. 10, no. 6, p. e0004795, 2016.
 9. G. Chowell, J. Cleaton and C. Viboud, "Elucidating Transmission Patterns From Internet Reports: Ebola and Middle East Respiratory Syndrome as Case Studies", *Journal of Infectious Diseases*, vol. 214, no. 4, pp. S421-S426, 2016.
 10. R. Welliver, J. Cherry, K. Boyer, J. Deseda-Tous, P. Krause, J. Dudley, R. Murray, W. Wingert, J. Champion and G. Freeman, "Sales of Nonprescription Cold Remedies: A Unique Method of Influenza Surveillance", *Pediatric Research*, vol. 13, no. 9, pp. 1015-1017, 1979.
 11. S. Magruder, "Evaluation of Over-the-Counter Pharmaceutical Sales As a Possible Early Warning Indicator of Human Disease", *Johns Hopkins APL Technical Digest*, vol. 24, no. 4, pp. 349-353, 2003.
 12. M. Pivette, J. Mueller, P. Crépey and A. Bar-Hen, "Drug sales data analysis for outbreak detection of infectious diseases: a systematic literature review", *BMC Infectious Diseases*, vol. 14, no. 1, 2014.
 13. J. Pastore, M. Zhao, A. Elangovan, "System and method for conducting real time active surveillance of disease outbreak", *US Patent App. 14/546,634*
 14. G. Wallstrom and W. Hogan, "Unsupervised clustering of over-the-counter healthcare products into product categories", *Journal of Biomedical Informatics*, vol. 40, no. 6, pp. 642-648, 2007.
 15. S. Bansal, G. Chowell, L. Simonsen, A. Vespignani and C. Viboud, "Big Data for Infectious Disease Surveillance and Modeling", *Journal of Infectious Diseases*, vol. 214, no. 4, pp. S375-S379, 2016.
 16. Report of the Havelock North Drinking Water Inquiry. Auckland, New Zealand, 2017.
 17. NZ Herald, http://www.nzherald.co.nz/hawkes-bay-today/news/article.cfm?c_id=1503462&objectid=11688495, 06/01/2018.
 18. Influenza Surveillance in New Zealand 2013, 2014, Institute of Environmental Science and Research Ltd (ESR): Wellington, New Zealand
 19. Influenza Surveillance in New Zealand 2014, 2015, Institute of Environmental Science and Research Ltd (ESR): Wellington, New Zealand