Science Diary

Introduction

This is the Diary I kept whilst I took upon the COVID-19 Wastewater modelling project. It's scrappy and meant to be a keepsake of notes from my individual work and from meetings.
The project began as a *Science Scholars* assignment, and was continued on over the summer (just for funsies).

Week 1

Had an hour long meeting with some people from ESR on Friday 22 July, Thomas Lumley, and some data modellers including one from Massey University. Discussed potential directions for each of us to take. Noted the following confounding factors that may be of interest in determining the number of COVID RNA bodies in wastewater:

- Virtual client network for weather data.
- Industrial waste.
- School term effects.
- New variant ~ more likely to go up?
- Large events?
- Size of RNA release for each variant.

Week 2

Had a meeting with Thomas Lumley on Tuesday 26 July. We discussed precisely what I should be working on.
This week I'll be investigating a couple of potentially confounding variables:

- Weather data
    - Clifro r package
    - Investigate whether we need one or multiple catchment sites to get a general idea for Auckland's rainfall.
        - Spaghetti plot the rainfall on each of these sites over time. (probably from the last year so we can get a good idea of how it changes in different sessions). Perhaps the season affects the catchment.
- Age data
    - Case data by age and location. ESR perhaps has this? No access at this stage. Thomas should be investigating.
- School term effects
    - Simple really
        - Superimpose vert lines representing school terms and holidays to attempt and observe upticks.
        - Remember that the introduction of a new variant is important information too. Don't want to confuse the two.
- Proportion of bodies in wastewater after events.
    - Same thing. Really more of just an exploration.

Week 3

I had another meeting with Thomas on Tuesday 2 July. Over the past week I had been working on what we had talked about, which included gathering weather, covid, and school term data and compiling graphics and models to generate insight. He made some suggestions to my code and I made amendments as he spoke. During the discussion with Thomas, he identified a few more investigation pathways I should take. These include the following:

First of all, we can assume that the population of Auckland has remained consistent over the past year.

Rain effects:

- Plot covid_bodies_per_person against rainfall.
- Ardmore Airport may not be recording rain data anymore. They're probably more interested in wind nowadays.
- The 'runoff' statistic may be more appropriate for the data. I should investigate whether this can be used in conjunction with regular rainfall, or in place of rainfall, and fit the model appropriately.
- It seems that when including the previous day's rain in the model, it has a more significant effect than the actual day's rain. How many days prior are important in predicting the number of bodies in water.
- Overall the effect of rain is small but may be significant (approximately 0 to 3.5%)
- Other regions, such as southern regions, with more severe rain and/or poorer infrastructure, may have a more significant effect of rain on the dilution of covid bodies in wastewater.
- Investigate how the proportion of bodies in wastewater changes with the introduction of rainwater. Perhaps this is a smart way of substituting the variables into the model (rain_avg or run_off_avg).

School term effects:

- Fit a model similar to the above, with a school term predictor.
- Fit a similar model without interaction between variant and school term, if significant, introduce interaction term.

Variant effect:

- Investigate the solo effect of a variant on the covid bodies per person
- Investigate ratio of genome copies per INFECTED person, and how it changes with the intro of a new variant.
    - Proportions of the variant prevalence in the community can be determined from the CGID file in drive. Or we can ask Mike Bunce if we can have the accurate proportions.

Week 4

Over the past week I had a look at all the highlighted green things from above. The most interesting thing I found was the strong significance in the number of SARS bodies per person in the population as predicted by the amount of rain. This tells us that it is very likely that Rain should be considered as a confounding factor in determining the number of cases in Auckland.
I held another weekly meeting with Thomas on Tuesday 9th of August. We discussed the findings and talked about next steps:

- Prepare findings in presentable format.
- Investigate school term effects.
- Investigate significance for some permutations of lagged days.
- Do similar analysis to other locations around New Zealand
    - Wellington
    - Christchurch
    - Dunedin
    - Palmerston North
    - Whangarei?

Week 5

I did not complete a lot of work over the past week. Although I did find even more significance in the Wellington and Christchurch areas when I ran the same analysis. I had an in person meeting with Thomas today, and Thomas thinks I am on to something. I am excited to keep looking into it. Over the next week I will continue the investigation this time keeping the following and previous in mind:

- Narrow selection of catchment centres down to wastewater local centres.
- Look into perhaps getting polygon areas for wastewater collection. These will give a better idea of what catchments will have more of an effect on the wastewater.
- Fit a higher-degree smoother for the date effect. A degree of freedom of about 12 would make sense for monthly change.
- Keep looking into making code more reproducible.
    - Use git data straight from R
    - Use *Clifro* from R

Week 6

Questions for Thomas:

- Can you explain why a quasipoisson model was suitable for this problem?
- Is my interpretation of the t-statistics and estimates correct?

Had a meeting. Minutes mentioned below:

Meeting of the SARS-CoV-2 in wastewater modelling team
25 August 2022: Commenced 1:00pm
In attendance: Bridget Armstrong, Michael Bunce, Richard Dean, Helen Morris, Alvaro Orsi, Andri Rachmadi (ESR)
Leighton Watson (University of Canterbury/Covid Modelling Aotearoa)
Toby Hayward, Thomas Lumley (Auckland University)
Apologies: Nigel French, Jonathan Marshall (Massey University), Lillian Lu (ESR)
Abbrev:
CMA = Covid Modelling Aotearoa
WW = wastewater
GC = genome copies
CSC = ESR - Christchurch Science Centre
KSC = ESR – Kenepuru Science Centre

No given agenda
Michael: Purpose of meeting:
1. Establish community of practice
2. Inflexion point modelling – is virus level in WW really increasing? How does that translate into disease burden? Percent chance of case rates increasing maybe hospitalisations?
3. Addition of inflexion point modelling to Obi Mobi dashboard https://www.esr.cri.nz/our-expertise/covid-19-response/covid19-insights/
4. WW trends modelling – correlation with cases or case rates
5. International comparisons

Leighton – asked to get Slack channel set up
Michael has set this up. Thanks! https://app.slack.com/client/T037WTUEHGV
Leighton demonstrated his data analysis based on weekly data from ESR publicly available GitHub https://github.com/ESR-NZ/covid_in_wastewater
Would like daily data to be available – this will be covered in contract/agreement between ESR and partners.
What data are available from other sources that are not utilising
Main takeaways from Leighton's analyses:

- Lower value of scaling factor is associated with what looks like a decrease reporting of cases (an increase in under-ascertainment)
- 5 week sliding window of moving averages were used to smooth weekly variations
  - It was a starting point as only weekly data were available at the time
- Assuming WW is a good proxy for case load, there is strong evidence that under ascertainment is increasing.

Thomas/Toby analysed how rain corresponds with GC per day per person.

Main takeaways:

- GC per day per person decreases with rainfall increase
- Rain measures are not necessarily from weather stations close to WW plants – There are some WW plants that have their own rain measurements – need find if these data are available from various WW.
- Check when WW processing moved to KSC – Bridget asked Jo Chapman and processing moved from CSC to KSC the week ending 22nd June 2022. There were some odd samples that were processed at CSC but all have since moved to KSC.

Action items:

Michael Bunce to advance the Schedule A of the academic contracts before the next meeting

Next meeting: Teams 1pm 8th September 2022. Click here to join the meeting

Meeting adjourned 2:05pm

Week Mid Semester

Another meeting with the covid wastewater modelling team. Had a presentation from Lillian Lu, who's presenting on behalf of someone else who is using neural networks to model case numbers using wastewater data.

Had to dash at 1:30PM for another job. Minutes below:

Meeting of the SARS-CoV-2 in wastewater modelling team
8 Sept 2022: Commenced 1:00pm
In attendance: Lillian Lu, Richard Dean, Helen Morris, Mike Bunce, Bridget Armstrong, Nigel French, Andri Rachmadi, Alvaro Orsi (later) (ESR)
Thomas Lumley, Toby Hayward (Auckland University)
Mike Plank (University of Canterbury)
Apologies: Leighton Watson (University of Canterbury & CMA), Jonathan Marshall (Massey University)
Abbrev:
CMA = Covid Modelling Aotearoa
WW = wastewater
GC = genome copies
RNN = recursive neural network
MASE = Mean Absolute Standard Error

No given agenda
Michael Bunce touched on Queenstown infectious diseases conference – lots of constructive/comments/etc with presentation about covid in WW.
Slack channel check in – all members are able to access slack.

Schedule A for academic contracts:
Document circulated as appropriate – for general WW projects
Mike Plank Happy
Thomas Lumley happy
Schedule A changes only, not the covid modelling part

Lillian Lu shared Alvaro Orsi's Notebook:
RNN model for inflexion point modelling Timeseries – past and future covariance.
The WW was the future covariance – everything else is current
Site predictions – some sites are good, some return terrible predictions. The model is trained using all time series form all sites. MASE for WW with cases is better than cases alone. How do we improve the sites that are going in the wrong direction? What about different sections of data (in time) to see how the predictions would work when we are coming in to times of large case surges.
Richard: Aside – high correlation of variants of can this be added to a model (or other models)?
Nigel - Are there patterns in the residuals that reveal more in the data – time correlation?
Mike Bunce mentioned the work by Thomas and Toby: Does this coincide with rainfall event?

Thomas – if you get a spike in Te Puke may not be because there are cases in Te Puke, - it may be because of an external shedder. The model predictions should depend on data from outside the watershed – infections aren't arising from nothing, they arise from another place and this needs to be integrated somehow.
Nigel Why do the weekly results correlate so well? WW = prevalence – cases = incidence which are not the same things
Mike Plank – Leighton are working on a semi-mechanistic model. Core of it: mechanistic info about how the virus spreading. Time lags affect what you see – reports and WW shedding detectability. How much lead time is there in the WW to cases? Is this sufficient to predict massive increase in incidence? How much case reporting has reduced? The overseas data is only about a 5 day lead – Mike Bunce to circulate paper. Missing waves because send infection may not be as intense as first infection therefore ppl are not testing/reporting
Mechanistic model – take in timeseries of cases and nowcast r0 and case load <<missed this please fill in>>??
Adding in WW may be give an idea that ascertainment rate may be dropping – with potential lags
Prevalence surveys – and is the reporting dropping by how much? How can we tell?
Mike Bunce mentioned : Leighton had a look at possibilities from CMA efforts (decrease in correlations over time)
Update Flow data to be forwarded to Toby and Thomas (Helen)
[1:34 pm] Andri Rachmadi Yes we do have PMMoV data and can try to normalize those with flow data (talking about population estimates)

Bridget: Scatterplot showing that the relationship between WW and case rate is changing at the national scale. For the past 5 weeks, incidence rate (case rate) has been slightly less than what was predicted by prevalence (GC/person/day).
Richard: Scatterplot for each site is similar - FYI neither of these take into account rainfall at all, or flow with weekly resolution.
Alvaro: work to be done: What is the interpretability of the models?
Richard: How do we share stuff in an appropriate way? Once a contract is signed, the data becomes more shareable with suitable Confidentiality clauses.
When we get [prevalence data =- level of confidentiality that is required for cases that have been surveyed. At the moment, hopes for SA2, but other co-variates may be required from cases data which may require tighter confidentiality/security of data.
Action items:
Mike Bunce to circulate paper in slack channel: How much case reporting has reduced? The overseas data is only about a 5 day lead
Helen to forward flow data to Toby and Thomas
Contracts to be read and signed
Meeting adjourned 1:45pm

Week 8

Mention to Thomas that I was given Variant Data for the Mid Semester test. I have reached out to Lisa Chen in hopes that they would point me in the right direction.

Map rain catchments to waste water catchments
Population estimates for each catchments
Create Visuals of rain catchments

Send Mackay List of wastewater sites of interest.

Using flow data see if it is

- useful at solo predicting wastewater samples genome
- Can rain predict daily flow
- Would including both make rain independent of dilution samples

Read about normalisation options for genome copies other than Flow

- PPmoV
    - Test if flow predicts PPmoV and vice versa
    - Perhaps both would be useful since PPmoV picks up population changes, and flow picks up more accurate dilution estimates.
- Paraxanthine (metabolised caffeine)
    - https://www.sciencedirect.com/science/article/pii/S0043135422009320

See what I can do with the current flow data and if it looks promising, reach out to Helen and see if we can use more data. Only useful if the data is within a few days to predict up ticks, or to have an advantage against case numbers.

Use the variant data to do another analysis.

Week 9

Use interpolation NIWA model for locations without rain data
https://niwa.co.nz/climate/research-projects/national-and-regional-climate-maps
For now, if a weather station is within 5 km of a waste catchment, we consider relevant
There were cases of catchments on islands (Mana Island intersecting Porirua) which may not be sensible.

I mapped and modelled catchment sites by their rain with varying estimates. No longer were the estimates all negative; there were definitely many sites with little data and/or significance in the estimates.

This week I also had my Science Scholars in-person presentation, where I displayed my current research. I was able to show my modelling estimates and give context to the research I am doing.

There is still a lot to work on in terms of correlating flow with rain. This is something I plan to work on next week.

Week 10

This week, I was preoccupied with other subjects, but I did some further work understanding the coefficients estimated by my models so far. A new understanding of the coefficient I measured is the beta is the additive effect of if the past 7 days had rained exactly the same amount. Beta to the power of the amount of rain over the past 7 days is the proportion of concentration of genome copies observed.

What is next is using a *Polynomial Distributed Lag Glm* model to model the effect instead. This will provide coefficients for the lag terms in order to give some consistency to the distribution of the rain impacting the flow. We expect the distribution of affected days to look right skewed as the day's lag increases.

We also had another meeting with the wastewater modelling group. I'm going to reach out to Lillian Lu to ask about their models. They seem to have used weather as a variable in their own model, so I am interested in how they quantified and measured that variable.

Week 11

This week I continued work looking into lagged GLM models. I managed to fit my first one but I am struggling to figure out how to use the coefficients to reproduce the graph of the estimated effects for each day.

- A *polynomial distributed lag glm* restricts the coefficients attached to the lagged terms to follow the distribution of a polynomial function. This makes natural sense in our case since we expect the effect of lagged rain days to follow a curved distribution.

I also had a look at the Rain effect on Flow. I estimated that every mm of average rainfall on a given day has somewhere between a 2-5% increase in flow. This makes sense because these numbers are roughly what we observed inversely in the proportions of concentrated genome copies.

Then I fitted the Flow data directly to the sars-genome copies data and observed the effect of the flow directly. Unfortunately, the flow data doesn't include many of the locations that we are interested in modelling (*for Auckland we really only have the "fringe" locations such as Omaha, Warkworth and Wellsford* ~ Thomas). I will be contacting Helen Morris for potential data on the other locations such as CBD Auckland, Wellington Moa Point, Rosedale North Shore etc.

Week 12

Talk to Bridget about potential error in the Rosedale Wastewater sample around July 2022.
Fit lagged rain and flow on the estimated number of samples.
Nigel French is getting pepper sales from countdown. No date on when that data is coming through.
ESR labs are measuring ppmov.

Week 13 (Beyond)

Alvaro Orsi model ML estimated measure for lag structure. Would be interested in comparing those estimates to mine.
When it comes to the meeting:

- Bring up Leighton's work. It doesn't seem to normalise the number of genome copies? Due to the significance of what I have shown, perhaps using flow to normalise the number of genome copies is a sensible idea and perhaps could have tighter confidence bands for the time being.
- Alvaro Orsi

I also created a visualisation and investigation of how much the variance can be explained by the transformed waste data given the rain estimates to much better results.

Week 14

Meeting with Thomas on Tuesday.
Objectives for the future:

- Will observing Flow be important in normalising the number of covid copies per case?
    - Estimate how much effect Flow has on the observed number of genome copies.
        - Quantify this effect. Poisson so percentage change.
        - Quantify for each region. As it stands, the sample location is a factor in the model. Will require extracting the coefficients and errors.
    - Pepper data. Nigel should be getting us access to the Countdown pepper data.
- Observing the effect of population movement.
    - Need population estimates and Wastewater catchments from Mackay
    - Need flow data from Helen Morris
    - PPMOV may also be of interest here also.

Also have a look at the effect of rain and flow on the number of genome copies.

Week 15

Going to have my next meeting with Thomas tomorrow, and I want to bring up the following queries:

1. The amount of Flow varies highly by region. I don't believe that the volumes of the water tanks are the same? Wouldn't this violate the current model since it assumes flow is independent of region?
2. Shouldn't we deal/investigate the high numbers of flow? Not all of the observations correlate with rain fall (reference figure in code block "rain on flow")

Completed more estimates and visualisations.

14/11: Did some tidying of the Documents

After my meeting with Thomas;

- Model is inappropriate, since we can use the flow data to explain directly the amount of dilution.
    - This is because copies-per-person is estimated by taking the yearly average of daily flow and multiplying it by the observed number of copies per litre of water.
    - Hence under the hypothesis that there is no effect of flow on the number of covid bodies, then the flow x copies per L should be about equal to the number of copies per person.
    - Copies_per_person is by the POPULATION not cases.

Next steps:

- Clean up this data.
- Create interesting visualisations of dilution.
- Talk to Mackay and Helen about Dunedin population and flow data.

Week 16/17

I have been in touch with Mackay and Helen.

- Helen is unable to get flow data for Dunedin. We'll need to figure out another means of getting this data.
- Mackay is able to provide population estimates, but not seasonal estimates.

Week 18

Had another meeting with the Wastewater group. I answered some questions on behalf of Thomas and myself, the group is interested in two things:

- How much variance is explained by Rainfall and/or Flow
- Whether we can trust Rainfall data instead of worrying about flow data.

Nigel is unlikely to receive the Pepper sales data.
Bridget Armstrong is developing Leighton's case ascertainment model in R; says she will upload her current work to GitHub for me to look at :)

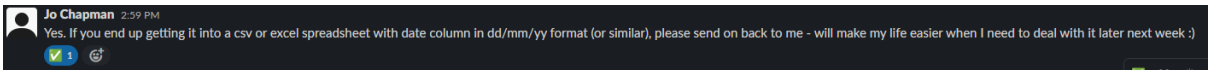
Alvaro is keen to share flow data from Dunedin. My plan is to run the same analysis (rain effect, flow effect, both in conjunction) on the Dunedin data. This should answer the two questions above, and develop our understanding of population movement affecting wastewater concentrations.

Week 19

Talking points

- Flow estimate for Dunedin
- Model copies per person in terms of rain
    - Add flow to model and see how much it improves the model
    - Extract coefficient and significance to explain improved accuracy.
    - Multiply by IQR to get “average” estimate of improvement when considered (when rain is held constant)
    - PPMOV data. Nigel has sales records, but I will need to contact Alvaro for ideas about where to find PPMOV wastewater data.
- Finishing up the project, uploading my data to a repo for the next student to use.

Immediately got to work and got the flow data for three waste catchment sites in Dunedin. I’d have to do some work with mapping the catchments using what Mackay has given me but that would be a job for another day.
Jo is also expecting me to tidy up the data and send it back to them sometime next week with dates corresponding to flow for computer reading purposes.



For now, I am tidying up my work on flow to bring to the next meeting.

Week 20

I've got to work on estimating the benefit of using flow over rainfall. The method I used to do this is by estimating the flow's significance and effect on the model when included alongside rainfall. What I found were very surprising and probably incorrect results and I wish to have them reviewed. I've emailed Thomas about it and hope he gets back to me before I have to present the findings on Thursday.

Would be good to model just flow against concentration.

Meeting notes

- From my work: Some locations sow significance between rain and concentration
    - AU_Rosedale
    - CA_Christchurch
    - WG_Porirua
    - WG_MoaPoint
    - WG_Seaview
- Flow provided more variance explained (by metric of significance) in the following locations:
    - AU_Rosedale
    - CA_Christchurch
    - WG_MoaPoint
- This tells me that at least in the City-Populations in Wellington and Christchurch, flow data is important. Although rain data isn't sufficient for most of Auckland.

Week 21

Created a GitHub repo with all the current wastewater files that I have been working on over the past 15 weeks. This will now be my current working directory for all changes that I will be making for the rest of this project.

Meeting notes

- Look at Dunedin
    - See if flow changes with time (population variance)
    - See if Concentration changes with time (population variance)
- Get PPMOV data from ESR
    - Joanne Hewitt has the data, see if ESR are ready to release it
    - If we can get it, have a set of normalisation approaches and see if PPMOV data gives a better normalisation over **rain, flow and no normalisation.**
    - Assume that consumption of peppers is constant. Any long-term seasonal variation would be explained in the smoother of the function.
        - Model the ratio of sars concentration to ppmov against rain and/or flow.
        - Model concentration of sars against PPMOV data.
            - It **should** be correlated; just a simple linear relationship.
        - Model concentration of sars against PPMOV with rain and/or flow.
    - REdo Flow file.
        - Add interpretations to the file.
        - P-value and AIC with both rain and flow.
        - Use confint instead of manually calculating confidence intervals.

Week 23? 2/1/23

Unsure about what week it is with respect to the first week of Sem 2.

Have had a look at the Dunedin data and completed an initial analysis of Rain on Conc, Flow on Conc and Rain/Flow on Conc.

- Rain and Flow both have significant effects on the concentration of genome copies in wastewater in the **Tahuna** region.
- Flow is shown to be necessary at estimating the effect.
- No significance is shown thus far to assume that Rain and/or Flow has an effect on the concentration of genome copies in **Green Island** or **Mosgiel**.

I've also tidied up the repository with understandable file names and .R source files that are useful for analysis.

Next steps are to figure out how to estimate the effect of population variance, and then incorporate that into the model to estimate the effect.

3/1/23

Still don't know where to find population fluctuation data. The best I can do is assume some proportion of the 21 thousand students who live in Dunedin leave over the Christmas period and then use that to try to determine a link.

10/1/23

Talk to Joan about getting PPMOV data. Thomas is convinced she has it.

Meeting notes

- Talk to Joan about getting PPMOV data.
- Try get some estimate for term dates for Otago University and see if it has an effect on:
    - Flow
    - Unscaled Genome Copies per Litre
    - Scaled Genome Copies per Litre (per person population).

Term dates in this respect are standing in for the information that would otherwise be explained by population variance.

- If term dates explain flow ~ we should be aware of what constant they predict.
    - It should also explain unscaled GC?

Overall I am unsure about the implications as much. Just do the analysis and come back to Thomas next week.

15/1/23

Did some population analysis. I feel like the smooth term to account for outbreaks is explaining the variance in population. I wonder how to work around this whilst still accounting for outbreaks? Perhaps another categorical term for known outbreaks instead?

16/1/23

Looking at the population with the adjusted flow.

- I.e. divide the concentration by the amount of measured flow and look for correlation.
- This could mean that population variance has an effect on the amount of flow.

More measure from Jo
Ignore the data from the beginning of sem 1 since there are other factors at play (lockdowns).

PPMOV data: Check if PPMOV is correlated with sars concentration.
See if the ratio of sars to PPMOV is predicted by rain or by flow.

Had another meeting with Thomas, above are just some of the notes that I took.

19/1/23

Got a message back from Jo Chapman

- PPMOV data is withheld by Joanne Hewitt and may not be accessible since ESR wants to analyse it first before making it available to us.
    - I've requested it anyway via slack.
- There are no new flow measurements for the past two months. The data we have is the most recent and available data.

Final Meeting 24/1/23

Had a final meeting with Thomas. Since we've been unable to get any data from Jo and Joanne thus far and the project has been somewhat overdue to pause up to and until the next student takes upon it, we've decided to stop here.
I have one final meeting with the group on Thursday and still intend on bringing up both the extra flow data and the PPMOV data with Jo and Joanne Hewitt.

- Tidy up the work so that it may be ready to present in the event that we will have to explain why the current flow data is insufficient.

26/01/23

Had my last meeting with the Wastewater modelling group today.
I am able to get access to flow data for Auckland over the holiday period, but not for Dunedin (Andri has been messaged).
A final report summarising all of the findings for this project should be made and uploaded to the GitHub repo for ease of reference.
Send a last email to Thomas for confirmation.

29th January 2023

The final report has been submitted to the git repo. It is bound to contain issues, but at least it is a single go to for the findings of the report. I've passed on a copy to Thomas for reading.

Thank you to Thomas Lumley for all the support over this project!

I really appreciate the time and effort you have spent on my first introduction to conducting research. You've been patient, kind and knowledgeable all the way through, and I couldn't be more pleased with the results of the investigation. I'm proud to have conducted thorough and supportive research to the Wastewater modelling group, and to have worked with such intelligent minds in the field. Thanks.