Validity of Queensland Wastewater Surveillance Program Methodology

Name: Brendan Wallace-Nash
Student Number: n9993304
Assignment: Insight report (Stream A)
Due Date:31/05/2021

Abstract
The insight report done at the request of Queensland health was to analyse the validity of the *Queensland wastewater surveillance program's* testing methods for SARS-CoV-2. Infrastructure, wastewater site density and travel routes were identified as possible factors in misrepresenting the rate of SARS-CoV-2 in catchment populations. In the analysis of the *Queensland wastewater surveillance program's* data, Queensland site data was split into north and south Queensland with south Queensland representing a region that had large Infrastructure, higher wastewater site density and more travel routes than that of north Queensland. The results of the analysis found that south Queensland had a much larger rate of SARS-CoV-2 detection than North Queensland, supporting the statement Infrastructure, wastewater site density and travel routes were misrepresenting SARS-CoV-2 detection. While it is suggested further research be made to support the analysis it has been recommended that the *Queensland wastewater surveillance program* should no longer be used as a metric of guidance for Queensland health's COVID-19 guidelines and restrictions as it jeopardises their reputations and could cause unnecessary harm to local economies.

This report is in response to *Queensland Health's* request for an analysis of the validity of their current *Queensland wastewater surveillance* program. In my analysis of the *Queensland wastewater surveillance program*, that detected SARS-CoV-2 in wastewater treatment sites, I proposed two questions for analysis. The first question being which wastewater treatment site on average had the largest detection of SARS-CoV-2, and secondly whether factors such as infrastructure and geological positioning of wastewater catchment proposed a bias in the rate of detection of SARS-CoV-2. For this report, I will be focusing more so on the second question I proposed and expanding on specific factors of infrastructure and geography that can impact wastewater surveillance results. The results of my analysis will be looked at through the context of the validity of the wastewater surveillance program and how it will impact stakeholders and future detection of COVID-19 and wastewater monitoring.

Wastewater Surveillance Program Analysis

infrastructure, wastewater site density and travel routes are the proposed factors of bias affecting the SARS-CoV-2 detection in *Queensland Wastewater Surveillance* data. For this analysis, infrastructure will be defined as hospitals, stadiums (above 10,000 capacity) and large shopping centres (more than 100 stores). Wastewater site density is a term used to describe the number of wastewater sites in an area. Travel route will be looked at as highways, train lines, airports, and shipping ports. These three factors form the foundation for the preformed analysis and reasoning as to why the current *Queensland wastewater surveillance program* is inaccurate.

The reasoning for choosing infrastructure, wastewater site density and travel routes is that these factors allow for non-residents of a catchment to contribute to the catchments testing results. In terms of infrastructure, hospitals, stadiums and shopping centres can attract large numbers of people from outside their wastewater sites catchment. In cases where there are SARS-CoV-2 outbreaks in neighbouring catchments, people from that catchment could use the infrastructure in neighbouring catchments and contribute to a positive SARS-CoV-2 wastewater result. While it is still important to monitor people's movement who have SARS-CoV-2 it is not the point of the *Queensland wastewater surveillance program*, as it is used to guide the Queensland government in making guidelines for identified COVID-19 hotspots. The same problem is present for areas with large travel routes, as places with highways and commercial transport may get wastewater samples from people that were only briefly stopping in the catchment and did not pose a risk of transmission but still contributed a positive sample. The density of wastewater site was identified as a problem factor because areas with many densely neighbouring wastewater sites can allow for the problems of infrastructure and travel routes to be more prevalent as people will be more likely to be moving between catchments.

To analyse the effects of infrastructure, wastewater site density and travel routes on the detection rate of SARS-CoV-2 in wastewater results the data was split into regions, north and south Queensland (Mackay used as boundary for north and south). The reasoning for splitting the data into these regions was done as south Queensland and north Queensland vary in term of infrastructure, wastewater site density and travel routes. South Queensland has a larger infrastructure than North Queensland to accommodate its much larger population. South Queensland also represents an area with a larger density of wastewater site catchments, having 25 in a relatively similar area of land to north Queensland who has 15 sites. South

Queensland also represents the majority of Queensland's biggest travel routes, having Queensland's two largest domestic and international airports and Queensland's largest shipping port, along with a busy highway system shared with New South Wales. Performing an analysis of the SARS-CoV-2 detection rate with data split into these two regions allowed for a representation of two regions that are drastically different in terms of infrastructure, wastewater catchment density and travel routes.

While splitting the wastewater site data into north and south Queensland allows for control of the identified problematic factors, there are still some uncontrolled factors in using this method. The main uncontrolled factor in the current split of data is the magnitude of unaccounted FIFO (Fly-In Fly-out) workers in north Queensland. North Queensland makes up the majority of Queensland's mining sector and for places like the Bowen Basin FIFO workers make up 19% of the population (Queensland Government Statistician's Office, 2021). This magnitude of FIFO workers presents a percentage of an unaccounted-for population as most mining camps reside outside the wastewater catchments and use septic tanks. This means that localised outbreaks to mining camps will not be represented in the North Queensland wastewater data. Septic tanks also present another unaccounted-for population as many rural residents' data would not be accounted for in wastewater sites if they use septic tanks. These factors should be considered as a contributor to the under-representation of the rate of SARS-CoV-2 in North Queensland wastewater data.

The data analysis focused on using the wastewater site location and detection data from *Queensland wastewater surveillance program*. Site data provided the name of the wastewater site, while detection was measured as detection or non-detection of SARS-CoV-2. It should be considered that there is no information provided on the threshold for SARS-CoV-2 detection. This means that detections rates could potentially be biased by the population that the wastewater site caters for, meaning smaller wastewater catchments sites could produce biased results If the concentration of SARS-CoV-2 was not controlled for in the wastewater testing. The testing rates of the *Queensland wastewater surveillance program* were not consistent among all sites so average test scores using mean was used to compare the sites. Splitting the data into north and south regions also created an uneven distribution of sites between the two regions. To account for this an overall average rate of detection was calculated for each region rather than comparing total site detection for each region. The result of the analysis process was a comparison of the average detection rate for south Queensland and north Queensland.

Analysis Results and Implications

As seen in *Figure 1* results from the analysis conducted found that south Queensland had a larger average rate of SARS-CoV-2 detection, measuring a positive result in 14.8% of all tests conducted, appose to north Queensland that only had a positive result of 8.1% of all tests. *Figure 2* also demonstrates that South Queensland wastewater sites also reported the highest positive test results, with Coombabah, Goodna, Nambour, Redcliffe and Fairfield having the largest positive test result rate in the state.

Figure 1: Average SARS-CoV-2 Positive Detection by Region

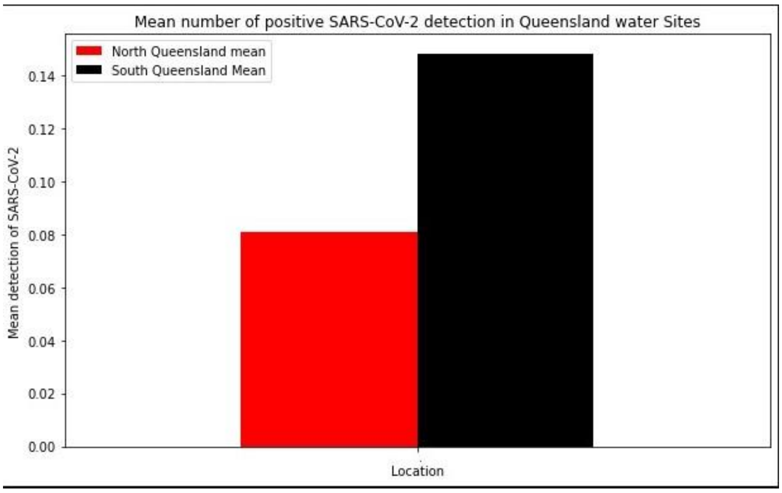
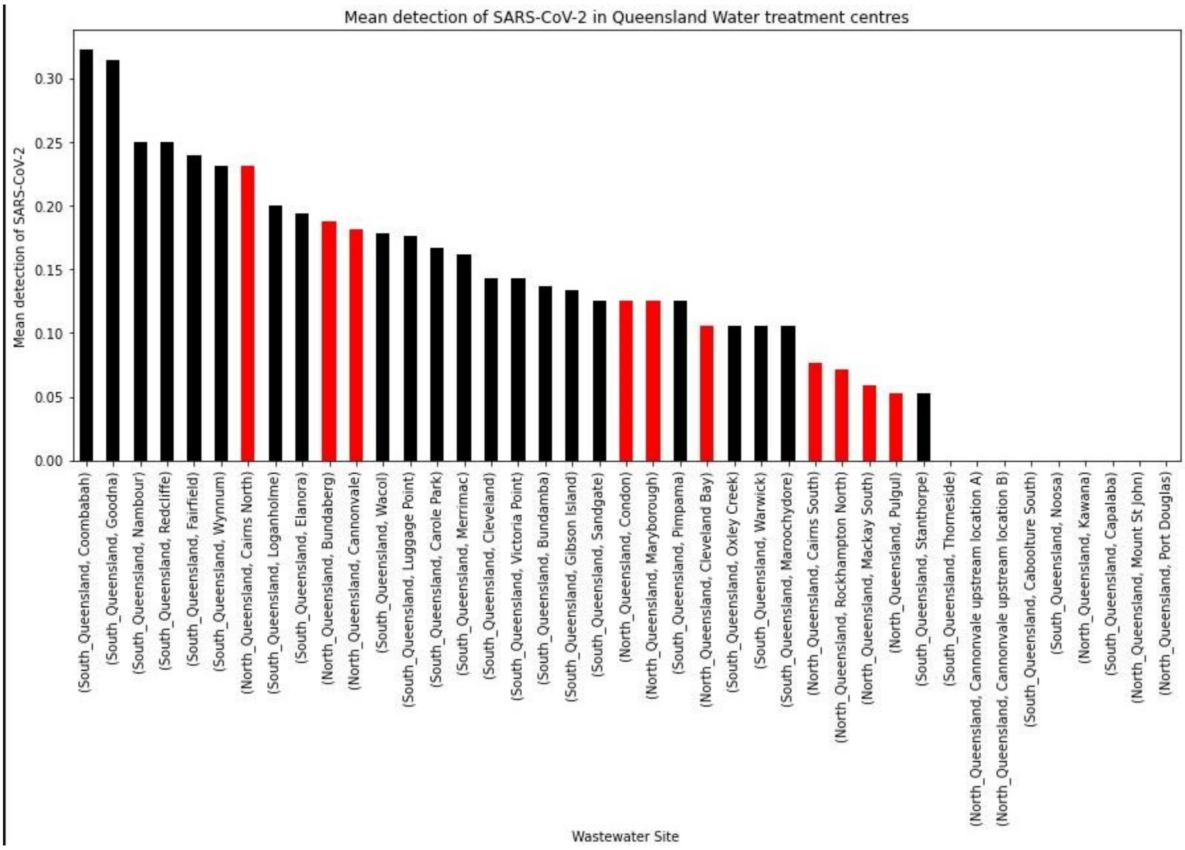


Figure 2: Average SARS-CoV-2 Positive Detection by Wastewater Site



When looking at the results of the data it can be implied that infrastructure, wastewater site density and travel routes influence the rate of SARS-CoV-2 detection, as south Queensland had a much larger detection rate than that of north Queensland. This claim is further backed up by the analyse of the Coombabah site, the wastewater site with the highest detection rate in all of Queensland. Coombabah had a positive test result for SARS-CoV-2 in 32.3% of its tests, a result well over south Queensland's average of 14.8%. When investigating the specifics of the Coombabah catchment area it is evident that infrastructure, wastewater site density and travel routes are all relevant factors to the Coombabah landscape.

The Coombabah catchment has many infrastructures that could have contributed to its large rate of SARS-CoV-2, with it being home to Harbour Town (one the largest shopping complexes in the state), Metricon Stadium (25,000 capacity) and the Gold Coast University Hospital. Coombabah's wastewater site is also within a 50km radius of two other wastewater catchments and has busy travel routes with both the M1, Gold Coast Highway and Southport airport positioned within its catchment.

When comparing the infrastructure, wastewater site density and travel routes of Coombabah to Port Douglas (wastewater site with the lowest SARS-CoV-2 detection, at 0%) the is a clear difference and indication as to why these factors are believed to bias test results. Port Douglas's wastewater catchment has no major infrastructure landmarks, with it being 20km away from the nearest hospital and has no significantly large shopping centres or stadiums. Port Douglas is also 60kms away from the next closest listed wastewaters site, and a further 267km away from its second-closest site. Port Douglas also features fewer travel routes with no airport and only one main highway. While there is a port in Port Douglas it is mainly used by tourism vessels and is not used for commercial shipping.

While the difference in SARS-CoV-2 detection rate in both south and north Queensland, and Coombabah and Port Douglas was equated to being a result of differences in infrastructure, wastewater site density and travel routes it should be noted there could be other mitigating factors. Both south Queensland and Coombabah have vastly larger populations than the regions they were compared to. If populations were not a controlling factor in wastewater testing it could be a reason why the results were so much higher. If population is found to be a factor in producing larger positive test results it still supports, the proposition that the current method used produces biased results for specific locations. It should also be noted that while inferences will be drawn from this analysis, conclusive conclusion of whether infrastructure, wastewater site density and travel routes effects in misrepresenting the presence of SARS-CoV-2 in catchment populations cannot be made until data from the current analysis is compared to that of data from SARS-CoV-2 clinical testing in those catchment regions.

The Effects of Invalid Testing on Stakeholders

The implications of the results of the analysis of *the Queensland wastewater program* highlights a potential problem with the validity of this metric of testing. While COVID-19's prominence and immediate threat have reduced in recent months, there is still cause for concern. The *Queensland Wastewater surveillance program* is still being used and it can be assumed that the current method would be used again to gather and interpret data in the case of another pandemic. The potential of an invalid testing method is a concern for the many stakeholders that use wastewater surveillance methods for guidance, and for the populations that reside within the catchments monitored.

While the general public is always the key stakeholder in terms of pandemic management the two key stakeholders that are considered most relevant to this analysis are Queensland Health and the local economies in the wastewater catchment area. While these are the two stakeholders identified in this report it should be considered that the results of this report have context to other states and countries that practice the same method of SARS-CoV-2 detection to build response plans for pandemics. Queensland Health, along with the University of Queensland and CSIRO, conducts the *Wastewater surveillance program* which is said to be used alongside clinical testing to develop responses to COVID-19. The use of invalid testing methods would not only affect the public perception of Queensland Health but would also lead to them developing responses that are not reflective of actual COVID-19 hotspot populations. Inaccurate COVID-19 response methods will also be damaging to the local economies that reside within the site catchment areas. Falsely identifying a catchment area as a COVID-19 hotspot would unnecessarily cause damage to the economy of that area as lockdowns or even cautioning the public of that area would massively reduce the flow of consumers to that area. As the *Queensland wastewater surveillance program* is still being conducted and data made publicly available local economies can still be affected by false results even if Queensland health does not enforce lockdowns or guidelines.

Can the Wastewater Data Still Be Used?

While the results of the analysis show that infrastructure, wastewater site density and travel routes do possibly misrepresent the catchment population results it should be considered that only areas with the presence of these factors are affected. This means that the current *Queensland wastewater surveillance program* could still be an effective SARS-CoV-2 monitoring method to guide Queensland health for catchments that do not meet the criteria of the misrepresented catchment.

To determine if a wastewater site catchment has the characteristics to be misrepresented or not, a Naive Bayes model could be used. The specific method of Naïve Bayes model used would be a Bernoulli Naïve Bayes model. The main assumption of this model is that the features (characteristics of wastewater catchments) used are independent of each other and are measured as binary (true or false). The target for a naïve Bayes in this context would be to predict if factors such as infrastructure, wastewater catchment density and travel routes have misrepresented the results of SARS-CoV-2 testing. The binary features for this model would be the presence of a single; hospital, stadium, shopping complex, train station, commercial airport, commercial shipping port and/or state highway. Due to this method having to use binary features catchment site density will be measured as true if there is a neighbouring wastewater site within 20km. Data of the listed features should be accessible to Queensland Health if they were to work in conjunction with the Queensland Department of Infrastructure. As all these features are represented as binary and all the chosen features are independent the assumptions of the Bernoulli Naïve Bayes model would be met.

The target used to train and test the accuracy of this model will be the results from the *Queensland wastewater surveillance program* in comparison to the clinical SARS-CoV-2 test results from site catchments at the same time. The target variable for a given catchment will be true (indicating the *Queensland wastewater surveillance program* is successfully representing SARS-CoV-2 presence in the population of the catchments) if wastewater test results are detected and SARS-CoV-2 positive cases in the area are above 0.01%, or if the wastewater test result is non-detected and the catchment population cases are below 0.01%. A negative target result (indicating the *Queensland wastewater surveillance program* is misrepresenting SARS-CoV-2 in the population of the catchments) is classified if the wastewater test result is detected and the catchments SARS-CoV-2 positive cases is below 0.01% or if the wastewater test results are non-detected and the catchments SARS-CoV-2 positive cases is above 0.01%. Using these metrics as the models results for training and testing accuracy will allow for the Bernoulli Naïve Bayes model to not only identify when a catchment is being overrepresented with SARS-CoV-2 results from wastewater testing but also underrepresented in the case where there is non-detection. This is relative to the context of rural places that could use irrigation systems or have large FIFO worker populations that would not be represented in the *Queensland wastewater surveillance program*).

Limitations and Ethical Concerns of Naïve Bayes Model

The risks of letting a machine learning model guide decisions should always be assessed critically. The naïve Bayes model that is to be used to classify if wastewater catchment testing results are valid would be responsible for the consequences of providing false guidance to Queensland health. The consequences of a wrong result guiding a decision of that magnitude could affect a region’s local economies and the likelihood of containing a SARS-CoV-2 outbreak.

The initial point of critique in the use of the naïve Bayes model is the collection of the data used in training and testing the model. If the data used to train and test the model is not accurate then the results of the implementation model will not be accurate. In the case of the comparison made of wastewater site detection results and actual diagnosed cases, having a population that was not diagnosed, falsely diagnosed, or even did not have access to testing would completely misrepresent a key factor in building the model. This would lead to a model that would incorrectly classify the water results as valid or invalid, providing false data to Queensland health that could again cause unnecessary risk to local economies affected by guidelines based on the false data.

There is also a potential for this model to be biased against lower socioeconomic areas. As the model is going to tend to identify areas with larger infrastructure and transport as areas with invalid testing, this could result in false-negative reports of invalid testing methods for higher socioeconomic areas and false-positive report for lower socioeconomic areas. This is based on the understanding that higher socioeconomic areas will have more vast infrastructure than that of lower socioeconomic areas. In the case of false-negative results for higher socioeconomic areas, there is not a significant concern as other metrics of testing will be used that will likely demonstrate an outbreak of SARS-CoV-2. However, in the case of false-positive report of valid testing in lower socioeconomic areas, the *Queensland wastewater surveillance program* will continue to be used and provide false guidance to Queensland health and affect lower socioeconomic areas local economies.

By allowing decisions to be made by a naïve Bayes model it opens the possibility for the model to be manipulated to report favourable results. In the case where Queensland health is in control of the model, there could be situations where the inputted data could be manipulated to produce favourable results. COVID-19 is a very political topic and the political landscape around lockdowns is very polarising. Allowing Queensland health to use a model in which they know how to manipulate to produce favourable outcomes to guide COVID-19 restrictions and legislation could be a potential issue.

Recommendations for Using Wastewater Surveillance

The last year has shown how important data collection has become for combating COVID-19. The effects of invalid data collection extends much further than the potential risk of an outbreak, as misinformed guidelines and regulations can jeopardise the reputation of authoritative bodies and damage local economies. The analysis of the *Queensland wastewater* *surveillance program* demonstrated a bias towards SARS-CoV-2 in regions with more prominent infrastructure, denser wastewater catchments and more frequent travel routes. While it cannot be said with certainty that the bias towards reporting higher SARS-CoV-2 detection in regions that meet this criterion is misrepresenting the population, there is a consistent trend that would suggest the listed factors would play a role in possible misrepresentation of catchment populations. While further analysis could be conducted to explicitly prove these findings correct, the potential risk of continuing to use the *Queensland wastewater surveillance program* in the meantime would be at the risk of local economies and Queensland health's reputation. If further research supports claims that the *Queensland wastewater surveillance program* is flawed in producing valid results for regions with large infrastructure, it should be noted that it can still be a valid method for other regions in Queensland. Adopting a naïve Bayes model to identify regions in which the *Queensland wastewater surveillance program* correctly represents SARS-CoV-2 in the population would allow for a greater depth of COVID-19 reporting for these areas. The limitations, bias and opening for manipulation in the use of this model should however be considered before its implementation. The insights in this analysis have been extensively critiqued and centred to the context of Queensland health and local economies. While these insights will be useful in future pandemics their implementation still need to be swift as COVID-19's re-emergence is unpredictable and identified stakeholders will bear the consequences of poor testing methods.