30-31 May 2022

Collaborative Workshop on Big Data Analysis of COVID-19

CSIR International Convention Centre

Hosted by SEPIMOD: This research team, which started at the onset of the COVID-19 pandemic in 2020, aims at spatial epidemiological modelling, which considers spatial covariates such as vulnerability, mobility, gender, age, comorbidities, and health care access, when assessing and fitting epidemiological models.

The SEPIMOD group consists of members across several institutions and disciplines and have been working on COVID-19 modelling in South Africa since the start of the pandemic. The team has expertise in statistical modelling, spatial analysis, GIS, programming, and epidemiology.

Details

The aim of this workshop is to bring together researchers, in South Africa and internationally, who have been involved in COVID-19 modelling to share research. There are several invited speakers already, namely Prof. Sheetal Silal, UCT and Prof. Bruce Mellado, WITS, as well as the research output of the SEPIMOD team.

The workshop will be presented in hybrid mode, but all speakers will present in person. If you would like to attend online please apply here. If you would to attend in person please apply here. In person attendance workshop fees are funded, thus selection will be competitive. Selected applicants need to cover their own travel and accommodation costs and will be expected to present their research at the workshop.

Deadlines: Thursday 27 May 2022 for online sign-up, Friday 20 May 2022 for in-person attendance

Enquiries: Inger Fabris-Rotelli

The workshop is funded by IDRC under the Africa-Canada Artificial Intelligence and Data Innovation Consortium.

PROGRAM OVERVIEW

Venue: CSIR Convention Centre, Pretoria

	MONDAY 30 MAY 2022	TUESDAY 31 MAY 2022
8:00 - 9:00	arrival tea and coffee	arrival tea and coffee
9:00 - 10:00	session 1	session 1
10:00 - 11:00		
11:00 - 11:30	tea and coffee	tea and coffee
11:30 - 12:30	session 2	session 2
12:30 - 13:30	lunch	lunch
13:30 - 15:00	session 3	session 3
15:00 - 15:30	tea and coffee	tea and coffee
15:30 - 16:30	session 4	session 4

COVID Screening

- All attendees to screen at this link before arrival at CSIR: https://rsvp.csir.co.za/index.php/476187?lang=en
- On the form, the event name can be inserted as: Collaborative workshop on big data analysis of COVID-19

Venue

- Our venue is the *Amethyst Room* at the CSIR ICC.
- Google Maps link to the ICC: https://goo.gl/maps/GmTgqPWvg1Qk21Tk8 (also see entrance to use below)
- Once inside, there will be directions to the room.

SPEAKERS and ABSTRACTS

Session 1: MONDAY 9:00 - 11:00

Sheetal Silal: Adaptive modelling for a novel virus: Reflections on COVID-19 in South Africa

SARS-CoV-2 is the most disruptive virus the world has faced in the last century. In the absence of previous experience and knowledge of the behaviour of this disease and the impact of measure to control, mathematical modelling and other analytical approaches have played a significant role in the global response to the pandemic. Due to the rapidly changing nature of the outbreak globally and in South Africa, mathematical models were updated regularly as new data became available. Changes in testing policy, contact tracing, and hospitalisation criteria all impacted the cases detected, treated and fatalities as well as the required budget for the COVID-19 response. This talk will follow the development of the National COVID-19 Epi Model (NCEM) from the start of the epidemic to date, reflecting on changes in assumptions and lessons learnt. The talk will end with considerations for modelling and policy on COVID-19 in the future.

Sally Archibald: Intermediate complexity spatial COVID models

XXX

Inger Fabris-Rotelli: A Spatial SEIR Model for COVID-19 in South Africa

The virus SARS-CoV-2 has resulted in numerous modelling approaches arising rapidly to understand the spread of the disease COVID-19 and to plan for future interventions. Herein, we present an SEIR model with a spatial spread component as well as four infectious compartments to account for the variety of symptom levels and transmission rate. The model takes into account the pattern of spatial vulnerability in South Africa through a vulnerability index that is based on socioeconomic and health susceptibility characteristics. Another spatially relevant factor in this context is level of mobility throughout. The thesis of this study is that without the contextual spatial spread modelling, the heterogeneity in COVID-19 prevalence in the South African setting would not be captured. The model is illustrated on South African COVID-19 case counts and hospitalisations.

Renate Thiede: Spatial variation in the basic reproduction number of COVID-19: A systematic review

COVID-19 has spread to more than 220 countries since it was first reported in Wuhan, China in December 2019. The basic reproduction number (R0), defined as the average number of secondary infected cases from one infected individual, is one of the most used epidemiological parameters in infectious disease modeling. The estimation of R0, however, is complex, especially when the epidemiology of the disease is not well known, as was the case in the early phases of the pandemic. R0 estimates from different countries ranged from 0.48 to 7.2 in the first months of the COVID-19 pandemic. The aim of this study was to investigate R0 variability globally and to investigate the spatial heterogeneity over the first half of 2020.

Inger Fabris-Rotelli: Modelling representative population mobility for COVID-19 spatial transmission in South Africa

The COVID-19 pandemic starting in the first half of 2020 has changed the lives of everyone across the world. Reduced mobility was essential due to it being the largest impact possible against the spread of the little understood SARS-CoV-2 virus. To understand the spread, a comprehension of human mobility patterns is needed. The use of mobility data in modelling is thus essential to capture the intrinsic spread through the population. It is necessary to determine to what extent mobility data sources convey the same

message of mobility within a region. This paper compares different mobility data sources by constructing spatial weight matrices at a variety of spatial resolutions and further compares the results through hierarchical clustering. We consider four methods for determining connectivity matrices representing mobility between spatial units, taking into account distance between spatial units as well as spatial covariates. This provides insight for the user into which data provides what type of information and in what situations a particular data source is most useful.

SESSION 2: MONDAY 11:30 - 12:30

Warren Bettenny: COVID-19 Wave Detection for District Municipalities in Gauteng

One of the most important and valued outcomes of epidemiological models is the ability to predict and pre-empt pending outbreaks or "waves" of a pandemic or epidemic. Since the start of COVID-19 there have been 4 distinct waves of infections in South Africa, each of which were accompanied by an increase in hospitalisations and deaths. It is for this reason that any foresight into the start of such a wave is invaluable to decision makers and public health officials when implementing plans to curtail the spread of the disease. This study investigates and implements the use of a wave detection algorithm developed by O'Brien and Clements (2021)1 within South Africa and – more specifically – the Gauteng region. The wave detection algorithm demonstrates both strengths and weaknesses when identifying impending waves. These are discussed along with the results of the study.

1: O'Brien Duncan A. and Clements Christopher F. 2021, Early warning signal reliability varies with COVID-19 waves, Biol. Lett. 17: 20210487.

Nada Abdelatif: Modelling the spread of COVID-19 in South Africa using stratified compartmental models

The novel coronavirus strand (SARS-CoV-2) first appeared in Wuhan, China in December 2019 and caused the respiratory syndrome COVID-19. A unique feature of COVID-19 is its non-uniform effect on populations. The effects of COVID-19 are more severe amongst the older and people with co-morbidities as seen by the higher mortality, infection and hospitalisation rates observed amongst these groups. This study models the spread of COVID-19 in South Africa March-August 2020 using stratified compartmental models to capture the population heterogeneity. An age and co-morbidity stratified compartmental model was built with additional compartments to capture the unique dynamics of COVID-19. A sensitivity analysis was performed to determine the models' sensitivity to start date and lockdown level to determine the optimal start date and to identify the effects of harsh lockdown restrictions on infections and hospitalisations. A parameter sensitivity analysis was also conducted to determine the parameters that needed to be re-estimated to improve model accuracy and to identify the age groups which were driving infections, hospitalisations, and deaths. These analyses showed that a prolonged harsh lockdown would have reduced infections by approximately 50% and delayed the infection peak by approximately 4 months. The analyses also showed that hospitalisations were driven by the 61-75 age group while infections and deaths were driven by the 76-90 age group. In addition, the model was most sensitive to infection duration, death rate and proportion of asymptomatic infection. These parameters were re-estimated to better capture the age and co-morbidity dependent dynamics of COVID-19.

Claudia Dresselhaus (online): A spatially explicit modelling strategy for Covid-19 predictions and wave risk analysis in Gauteng

Countries around the world implemented their national vaccination campaigns to gain control of the COVID-19 pandemic that has torn across the globe in 2020. In this context, a nuanced spatial SEIRDV model was proposed to incorporate the immunity level of the population to predict the occurrence and spread of COVID-19. The spatial SEIRDV approach is able to model the infection counts of the COVID-19 disease, by taking into account spatial areas and its characteristics such as its vaccination coverage during a given

time period. This may assess the extend to which prevention mechanisms such as to what extend vaccinations minimise the severe burden of the COVID-19 pandemic and if a particular spatial area drives COVID19 infections or not using a spatially explicit model. The conclusion of this study is that the number of immune people that recovered in asymptomatic and mild infections, the longevity of patients being sick of infected as well as the number of people that are immune towards COVID-19 are the main drivers of the spread of COVID-19. This could advise Gauteng's Department of Health in terms of their pro-active immunisation policies.

Marié Vogel-Jacobs: A high resolution human movement model using small area estimation

Spatial movement models are important for the improvement of epidemiological models of contagious diseases, such as covid-19. High resolution models are needed for accurate predictive capabilities, but there is often a lack of sufficiently detailed data, or access to it. Small area estimation may come to the rescue, by using higher resolution covariate data together with lower resolution mobile network data, to predict a more accurate picture of human movement patterns.

Gandhi Jafta: Spatial-temporal topic modelling of Covid-19 related tweets in South Africa

In a society as socially and economically diverse as South Africa, a uniform response to crises can often be found wanting since it does not cater to all its recipients' lived realities. The Covid-19 pandemic highlighted this issue. A case in point is the #whichSouthAfricans trend that started on Twitter. The trend began when a political leader – in his critical response to government regulations - claimed to be speaking on behalf of South Africans.

Topic modelling is a popular natural language processing technique that is used for finding hidden topics in documents such as tweets. Tweets contain useful information about public opinion and the lived realities of its users. We propose that using topic modelling can allow us to extract this information. Additionally, adding a spatial dimension allows us to see where the topics from the tweets are coming from. This can aid in directed response to different areas. Covid-19 response is dynamic and changes over time, it is for this reason that a temporal dimension will be added.

Ogone Motlogeloa: Assessing the role of climatic conditions on temporal seasonality of lower respiratory diseases in South Africa

It is estimated that approximately 47,000 cases episodes of influenza-associated severe acute respiratory illness (SARI) and approximately 9,500 influenza-associated all-cause deaths occur annually in South Africa. Internationally there has been a distinct seasonality in caseloads of lower respiratory disease, and this has been found to be associated with seasonal changes in climate. This study explores the seasonality of lower respiratory disease cases, and the role of climate, in South Africa This study explores the seasonality through three approaches. The first involves an analysis of recorded positive test outcomes for influenza, RSV and four coronaviruses at Baragwanath Hospital over a 10-year period. The second involves the analysis of medical aid claims for a broad basket of respiratory diseases across South Africa, again for a 10-year period. The third involves an analysis of the perceptions of healthcare workers and the public pertaining to the seasonality of respiratory diseases, and the role of climate. Understanding contemporary dynamics in spatio-temporal patterns of disease incidence and the climate connection are imperative in being able to track, and adapt to the impacts of climate change.

Session 3: MONDAY 13:30 - 15:00

Siphiwe Thwala (with Joan Byamugisha, Richard Young, Tamara Govindasamy online): A Semantic-based Approach to Sentiment Analysis

Sentiment analysis is a common task in Natural Language Processing (NLP). The approaches taken to perform this task range from rule-based syntactic methods, to lexical methods, up to corpus-based methods based on machine learning methods. Standard sentiment analysis libraries have been built from lexical methods, while corpus-based methods are regarded as the state-of-the-art. We carried out an investigation to evaluate how well lexical-based and corpus-based standard libraries handle edge cases of sentiment analysis, those cases that are representative of logical complexity of human language. Based on the results of this investigation, we developed an approach to sentiment analysis, which is based on the sentiment of the lexical items within an input text. Our approach follows the definition of "semantic as "Relating to meaning in language"', and we relied on lexical parsing using the Answer Constrained Engine (ACE) as well as logical representations of knowledge in WordNet. We evaluated our approaches on the same edge cases that were used to evaluate lexical-based and corpus-based standard libraries, and the results show that our method out-performs these two methods in identifying sentiment correctly.

Gillian Maree: Data and decision-making: The COVID-19 pandemic in the Gauteng City Region

The GCRO has worked with the Gauteng Provincial Government and key role-players during the pandemic to provide research, data and insights to help inform response planning to COVID-19. A range of data sets from our own Quality of Life Survey, to open source imagery and geocoded COVID-19 case data were analysed and modelled to provide insight into the spatial patterns, impacts and trends on the Gauteng City-Region. We found that that visualisation can be an ideal tool for communicating insights from large and complex datasets with those who needed to make policy or make life-changing decisions quickly. Data, and how it is visualised and made available, has had an impact in shaping policy during the pandemic. However, there are also limitations – both from the data itself and with the diverse range of demands and priorities that decision-makers face. This presentation will show how the GCRO used a diverse range of data in our work on the pandemic to visualise its impact on the City-Region and support response planning.

Rhena Delport: Complexities of data collection in informal settlements

XXX

Session 4: MONDAY 15:30 - 16:30

Neville Sweijd: Consideration of Environmental Factors in COVID and other Infectious diseases

Here we will review what we know about the role of climate and environmental drivers of C19 in the pandemic and its role in driving endemic infectious diseases. We will also present some key considerations in the relationship between climate variability and climate change on infectious diseases and in human health in general and how this should be integrated in disease modelling. Here we will consider a range of Climate Sensitive Diseases and Disorders and present some information about how this can be more formally integrated into ongoing climate services governance.

Jude Kong (online): The impact of social, economic, and environmental factors on the dynamics of COVID-19, Estimation of epidemiological parameters and ascertainment rate from early transmission of COVID-19 across Africa

The COVID-19 pandemic has reached a stage where there is now sufficient data to infer whether the basic reproduction number (R0) varies across countries, and what demographic, social, and environmental factors, other than interventions, characterize vulnerability to the virus. In this talk, I will present the first global estimate of R0 across all continents, and the results of a comprehensive investigation on what social, economic, and environmental factors characterize vulnerability to the virus. Understanding how space and time-dependent factors predispose a community to a different COVID-19 rate of increase is essential to assessing the efficacy of interventions.

Session 1: TUESDAY 9:00 - 11:00

Bruce Mellado: The role of AI in managing the COVID-19 pandemic

In this presentation a number of projects and showcases will be succinctly described regarding the use of Artificial Intelligence in the management of the COVID-19 pandemic. This includes, the implementation on Machine Learning in the detection of hot-spots, development of early detection algorithms, vaccination strategies, dissecting the problem of vaccination hesitancy, monitoring social distancing, among others. The use of AI for pandemic preparedness in general will also be discussed and showcases will be provided.

Meghan Malaatjie (online): The use of machine learning to identify gender differential outcomes in a South African multi-center COVID-19 cohort

Women have been inordinately affected by the COVID-19 pandemic globally, in comparison to their male counterparts. Previous disease outbreaks such as the Zika virus and Ebola outbreak in South America and West Africa respectively, have shown that gendered norms experienced during disease outbreaks disproportionately affect women and place them at risk of poor health outcomes. This study aims to investigate whether Machine Learning techniques can provide supplementary information of gender differentials in COVID-19 hospitalizations and hospitalization outcomes throughout the four waves of the pandemic, in the Gauteng province of South Africa. A deep neural network (DNN) was trained and used to separate two classes of data sets: severe case of disease (mortality, intensive care, and high-level care) and less severe case of disease, dis-aggregated by gender. Observed differences in COVID-19 hospitalization between men and women were more significant in the 20 – 40 year age group with a COVID-19 hospitalisation ratio of 1:3 for males to females. This large difference in hospitalization frequency was observed for less severe cases of disease.

Nicholas Perikli: A Natural Language Processing approach to Probe COVID-19 Vaccine Hesitancy from Tweets in South Africa

This paper is targeted at an NLP venue, and as such, it focuses on presenting the main contribution through the NLP methods, with vaccine hesitancy as a use case, in which a dataset of 50 000 tweets from RSA were extracted and hand-labelled into one of three sentiments - positive, negative, neutral. The machine learning models used were LSTM, bi-LSTM, SVM and RoBERTa base, whereby their hyperparameters were carefully chosen and tuned using the WandB platform. All models were found to have have a value accuracy above 80%, with RoBERTa achieving an impressive 84% accuracy.
Pre-processing methods involved both Semantic-based and Corpus-based approaches, respectively. It was found that both methods were equally competent in training the models. An LDA was then performed on the mis-classified tweets and from the results we were able to to draw conclusions on how to further improve the accuracy of these models.

Finn Stevenson: Development of a RNN LSTM based Risk Index for identification of additional COVID-19 case waves

Our experience since the beginning of the COVID-19 pandemic has demonstrated the need for case data prediction models as well as alert systems for additional waves of COVID-19. The availability of a variety of newly developed indicators allows for the exploration of multi-feature prediction models for case data. This research documents the development of an AI-powered additional COVID-19 case wave alert system for South African provinces. Newly released indicators of human mobility, government policy stringency and COVID-19 epidemiological parameters are used as inputs to the early alert system. The model used for the system is a Recurrent Neural Network (RNN) with Long Short Term Memory (LSTM), an architecture known for its applicability to multi-time-step multivariate prediction problems. The early alert/detection system functions by predicting future daily confirmed cases based on a series of features that include mobility and stringency indices, and epidemiological parameters. The model is trained on the intermittent period in between concurrent waves, in all of the South African provinces. The early alert system has been used by the Gauteng provincial government for surveillance of additional waves. Demonstrations of the system's use for identification of the 4th and 5th waves will be shown.

Session 2: TUESDAY 11:30 - 12:30

Samuel Manda: Development and validation of a prediction model of COVID-19 hospitalisation and mortality using a South African private health insured population

A prediction score for COVID-19 hospitalization or mortality has the potential to aid clinicians in the development of better risk mitigation and improvement of quality of care. and avoid potential hospital admissions. Using a large cohort of COVID-19 patients from a South African private health insured population, we developed and validated a prediction model of COVID-19 hospitalization and mortality. Multivariate logistic regression was used to identify the independent risk factors related to either COVID-19 hospitalisation or mortality. The area under the receiver operating characteristic curve (AUROC) and predictive accuracy were used to evaluate the model's discriminative ability. A total of 188,292 members who tested COVID-19 positive over the period 1 March 2020 – 28 February 2021 were extracted. Overall hospitalization and mortality risks were 18.8% and 3.3% respectively. Higher hospitalisation and mortality were observed ed in patients aged at least 65years, had at least 3 comorbidities, and were males. Our prediction models achieved good predictive accuracies and AUROCs. Our prediction risk scores can predict the risk of the disease COVID-19 hospitalisation and mortality.

Tarylee Reddy (online): Statistical approaches for estimating vaccine effectiveness in a real world setting

The Sisonke trial, was a Phase 3b implementation trial conducted between February and May 2021, involved vaccination of approximately 500000 healthcare workers in South Africa. The analysis of vaccine effectiveness involved the integration of several data sources including: medical scheme data, the national death registry, NMC and DATCOV data, and the EVDS. In this presentation we describe the computationally intensive approach of a simulated clinical trial using a matched cohort design. We also discuss current challenges in the evaluation of vaccine effectiveness including: assessing durability of vaccines, effectiveness of booster doses, other study designs and the availability of data.

Michelle de Klerk: Spatial modelling of COVID-19 lineage in South Africa

In the work we present a proposal on how to approach the spatial modelling and tracking of COVID-10 lineage in South Africa.

Session 3: TUESDAY 13:30 - 15:00

Hui Zhang: Statistical methods and considerations for COVID-19 vaccine evaluations

The COVID-19 pandemic has lasted for over two years and caused immense economic, medical, and social devastation to the whole world. As of May 2022, there have been over 500 million COVID-19 cases detected worldwide, and over 6 million related deaths. To fight with COVID-19, multiple vaccines have been developed. Therefore, evaluating the efficacy of COVID-19 vaccine has been a critical scientific question, for which biostatistics plays an important role. In this workshop, we will discuss the statistical design and analysis of clinical trials to evaluate COVID-19 vaccine under controlled experiments, as well as the statistical modelling of post-market surveillance by real world COVID-19 vaccine studies.

Zhengjun Zhang: The Existence of at Least Three Genomic Signature Patterns and at Least Seven Subtypes of COVID-19 and the End of the Disease

Hoping to find genomic clues linked to COVID-19 and end the pandemic has driven scientists' tremendous efforts to try all kinds of research. Signs of progress have been achieved but are still limited. This paper intends to prove the existence of at least three genomic signature patterns and at least seven subtypes of COVID-19 driven by five critical genes (the smallest subset of genes) using three blood-sampled datasets. These signatures and subtypes provide crucial genomic information in COVID-19 diagnosis (including ICU patients), research focuses, and treatment methods. Unlike existing approaches focused on gene fold-changes and pathways, gene-gene nonlinear and competing interactions are the driving forces in finding the signature patterns and subtypes. Furthermore, the method leads to high accuracy with hospitalized patients, showing biological and mathematical equivalences between COVID-19 status and the signature patterns and a methodological advantage over other methods that cannot lead to high accuracy. As a result, as new biomarkers, the new find-ings and genomic clues can be much more informative than other findings for interpreting biological mechanisms, developing the second (third) generation of vaccines, antiviral drugs, and treatment methods, and eventually bringing new hopes of an end to the pandemic.

Kathryn Arnold: Creating a Set of High-Resolution Vulnerability Indicators to Support the Disaster Management Response to the COVID-19 Pandemic in South Africa

This talk presents the "COVID-19 Vulnerability Dashboard" for South Africa, developed by the CSIR for the National Disaster Management Centre (NDMC). It maps vulnerability to COVID-19 for the whole of South Africa, down to the level of the 103576 enumerator areas (EAs). The COVID-19 Vulnerability Dashboard was aimed at helping the NDMC, local authorities and other stakeholders with disaster risk reduction (DRR) and evidence-based decision making during the early stages of COVID-19 response in March 2020. Several national government departments have also used the Dashboard for planning support. South Africa has large populations around the country vulnerable to COVID-19 because of the triple challenges of poverty, inequality and employment, and the high levels of HIV/AIDS and tuberculosis; high potential for rapid spread because of many dense informal settlements; and limited health resources. The COVID-19 Vulnerability Dashboard drew on CSIR data and expertise in spatial analysis and disaster risk reduction of human settlements. Using a multi-criteria analysis approach, a set of vulnerability indicators based on domain knowledge were created, which was peer-reviewed by expert groups. These were disseminated by dynamic spatial mapping through an interactive, online dashboard.

Session 4: TUESDAY 15:30 - 16:30

Mogesh Naidoo and Juanette John (online): Modeling impacts of COVID-19 lockdowns on emissions and air quality within the Highveld region

The COVID-19 lockdowns during 2020 presented an opportunity for a global air quality experiment. Ground based and satellite measurements revealed significant improvements to air quality during the

lockdowns associated with the reduction in emissions from key sources in urban areas such as traffic and industry. Thus, what is generally investigated theoretically during air quality management studies has occurred in a real-world context. This global experiment made clear that drastic emission reductions led to cleaner air and potentially better health outcomes. However, in many developing countries, impacts on air quality were not straight forward, as some emission sectors such as residential fuel combustion experienced increases and have relevance to human health. Additionally, non-linear atmospheric chemistry has the potential to increase secondary pollutants such as ground level ozone. Researchers within the CSIR's Climate and Air Quality Modelling (CAQM) group are collaborating with those at University of London, University of Pretoria and North-West University to investigate the dynamics and drivers of the response of air quality to COVID-19 lockdowns through multiple data streams including ground level measurements, satellite retrievals and chemical transport modelling. The modelling aids investigation on a more spatially comprehensive extent than the measurements, while also providing a further means to understand what would have happened if no lockdowns were imposed or if individual emission sectors behaved differently.