**COVID-19 DATA ANALYSIS RESEARCH**

**REGINA EDITH ARTHUR**

**ADVANCE DATA ANALYSIS**

**Topic Modelling**

**Dataset**

The dataset selected for this topic modelling project consists of over 179,000 tweets related to COVID-19, collected from Twitter using the #covid19 hashtag.

This dataset, sourced from Kaggle and compiled by Gabriel Preda through the Twitter API, captures a wide spectrum of public discourse during the pandemic.

The text data includes short, informal messages that reflect real-time reactions, news dissemination, and personal experiences, making useful for topic modelling applications.

Key features of the dataset include:

* A large volume of text entries with no missing values in the primary text column
* A temporal range starting from July 25, 2020
* Metadata such as user location, date, and hashtags that could support extended analyses.

3. Data Pre-processing

Data pre-processing was performed in R using base functions and packages such as textmineR, tm, and textclean. Key pre-processing steps included:

1. Lowercasing all text
2. Removal of:

* URLs
* Punctuation and special characters
* Stopwords (using a custom and built-in list)
* Words with less than three characters
* Tokenization and lemmatization
* Construction of a document-term matrix (DTM)

After pre-processing, the DTM consisted of 2,000 documents (tweets) and 3,442 unique terms.

**Topic Modelling with LDA**

Using the textmineR package, Latent Dirichlet Allocation (LDA) was applied to the pre-processed DTM to identify latent topics. Multiple LDA models were trained with topic numbers k = 2, 4, 6, and 8 to evaluate model quality based on coherence scores.

**Topic Coherence Evaluation**

The model with six topics (k = 6) was selected as optimal due to its relatively high coherence, suggesting more interpretable and semantically meaningful topics.

**Topic Interpretation**

The top terms for each of the six topics are as follows (manually labelled based on semantic content):

Topic 1: Public Health and Safety

Terms: mask, wear, safety, guideline, distance, protect

Topic 2: Political and Social Commentary

Terms: trump, government, response, policy, america, blame

Topic 3: Medical Updates and Vaccines

Terms: vaccine, pfizer, approved, trial, dose, effective

Topic 4: Economic Impact

Terms: economy, job, lockdown, closed, unemployment, support

Topic 5: Global Statistics and News

Terms: case, death, report, number, update, global

Topic 6: Personal Reactions and Sentiment

Terms: scared, stay, home, hope, pray, tired

The dominant topic per document was identified, enabling classification of tweets based on their most prominent theme.

**Word Frequency Analysis**

The most frequently occurring terms across all tweets included:

* covid
* coronavirus
* mask
* lockdown
* vaccine

These frequencies aligned well with the key terms in the LDA topics, validating the coherence of the model.

**Conclusion**

The six identified topics captured diverse aspects of the pandemic discourse, including public health, policy, medical progress, and individual sentiment.

**Directed and Undirected Graphs**

**Dataset**

The dataset selected tracks confirmed cases, reported deaths, and recoveries from COVID-19 across countries. It is sourced from reputable organizations like the World Health Organization (WHO), European Centre for Disease Prevention and Control (ECDC), and compiled by Johns Hopkins University CSSE.

Given the strong relationships between infection rates, recoveries, and deaths, this dataset is ideal for applying graphical models to explore dependencies.  
It contains clean, normalized, country-level time series data in CSV format, focusing on three key variables:

* Confirmed cases
* Deaths
* Recovered cases

The structured and interrelated nature of the data makes it well-suited for both directed (Bayesian Networks) and undirected (Markov Random Fields) graphical analysis.

(Link : <https://datahub.io/core/covid-19#readme> )

**Data Preprocessing and Exploratory Analysis**

The dataset was first examined for missing values and inconsistencies, with rows containing NA values in key columns—Confirmed, Recovered, and Deaths—systematically removed.

Descriptive statistics provided insight into the distribution of case counts across countries. A correlation heatmap was then constructed, revealing a strong positive correlation (0.92) between confirmed cases and deaths, and a weaker correlation (0.28–0.32) with recovered cases, suggesting distinct underlying dynamics between these variables.

**Graphical Model Implementation**

For model implementation, a directed graphical model in the form of a Bayesian Network was constructed using the bnlearn package. The structure was learned through a hill-climbing algorithm based on the cleaned numeric dataset. The resulting network indicated direct dependencies between confirmed cases and deaths, reflecting a probabilistic relationship where high confirmed counts significantly increase the likelihood of elevated deaths.

Inference was performed via conditional probability querying, where the probability of deaths exceeding 1000 given more than 50,000 confirmed cases was approximately 0.90, underscoring the predictive strength of the model.

An undirected graphical model was then derived by moralizing the Bayesian Network, resulting in a Markov Random Field (MRF) structure. The MRF graph, plotted using the igraph package, preserved key dependencies while removing directional assumptions. This undirected structure offered a symmetric view of variable relationships, suitable for tasks where causal inference is less critical but joint distribution properties are of interest.

**Inference, Interpretation, and Conclusion**

Inference and interpretation highlighted meaningful statistical associations and dependencies within the COVID-19 data. The directed model emphasized conditional probabilities and causally plausible paths, while the undirected model supported structural analysis of co-dependencies.

These models demonstrated real-world applicability by offering interpretable frameworks for assessing the impact of case surges on fatalities and recoveries.

Conclusions drawn from both model types were logically derived, showcasing the utility of graphical models in epidemiological analysis and contributing valuable insights into the progression and outcomes of the COVID-19 pandemic.

**Time Series Modelling**

**Dataset**

A time series dataset obtained from the World Health Organization (WHO), titled *WHO-COVID-19-global-data.csv*, was used in this analysis.

The dataset contains daily records of reported COVID-19 cases and deaths from countries around the world, covering the period from January 3, 2020, to April 2025. The variable of interest is the global daily new cases, recorded under the column *New\_cases*. The data is collected and reported at a daily frequency, making it appropriate for time series modelling.

(Link: <https://data.who.int/dashboards/covid19/data?n=c> )

**Data Pre-processing**

A comprehensive analysis of global daily new COVID-19 cases was conducted using time series modelling techniques. The raw dataset, sourced from the World Health Organization, was pre-processed to correct negative values and address missing entries, which were replaced with zero due to data correction artifacts. Aggregation at the global level was performed, and a consistent date format was applied for time series construction.

**Exploratory Data Analysis (EDA) and Stationarity Assessment**

Initial visualization of the time series revealed pronounced variability and irregular trends in case counts. Stationarity was assessed using the Augmented Dickey-Fuller (ADF), Phillips-Perron (PP), and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests. While ADF and PP indicated stationarity, the KPSS test contradicted this, suggesting non-stationarity. First-order differencing was applied, after which all tests confirmed stationarity, validating the use of differenced data for modelling.

**Model Selection and Fitting**

ACF and PACF plots informed the selection of appropriate model parameters, and the auto.arima() function proposed an ARIMA(5,0,5) model. The model was fit directly to the original series without further differencing, incorporating both autoregressive and moving average terms. The model's coefficients were statistically significant, and diagnostic checks using the Ljung-Box test revealed significant residual autocorrelation, indicating that although the model captured broad patterns, residual structure remained.

**Forecasting**

A one-year forecast was generated from the final model. The projections revealed high uncertainty with wide confidence intervals, underscoring the volatile nature of pandemic dynamics. Despite this, the model captured broad seasonal patterns and can inform general trend expectations.

**Conclusion**

Findings support the applicability of ARIMA models in COVID-19 case forecasting, but high forecast variance suggests that results are best interpreted with caution. Policymakers may use the projected trend envelope to gauge readiness strategies rather than absolute counts. Visualizations such as time series plots, ACF/PACF charts, and forecast intervals were integral in conveying key insights throughout the analysis.

**Spatial Modelling**

**Datasets**

For the spatial modelling component of this project, three key geospatial datasets were selected from reputable sources: a subnational COVID-19 dataset, a regional population dataset, and a Ghana administrative boundary shapefile.

The COVID-19 dataset (*gha\_subnational\_covid19\_hera.csv*) was sourced from the Humanitarian Data Exchange (HDX) (<https://data.humdata.org/>) and provides region-level data on confirmed cases, deaths, and gender-disaggregated statistics from March 12, 2020, to May 19, 2021.

The population dataset (*Ghana Administrative Population Estimates 2021*) is also from HDX, contributed by UNFPA, and offers 2021 regional population estimates essential for normalizing case data and enabling meaningful comparisons.

Lastly, the Ghana ADM1 shapefile from GeoBoundaries (geoBoundaries-GHA-ADM1\_simplified.shp) (<https://www.geoboundaries.org/> ) defines the geographic boundaries of Ghana’s 16 administrative regions.

These datasets are well-suited for spatial modelling due to their geographic resolution, compatibility, and reliable provenance, making them ideal for analysing spatial patterns and dependencies in COVID-19 outcomes across Ghana.

**Exploratory Spatial Data Analysis (ESDA):**

Data pre-processing involved standardizing column names, removing entries with undefined regions, and ensuring consistency across spatial datasets. Coordinate reference systems were verified for projection consistency between shapefiles and attribute data.

A summary dataset was created by aggregating total confirmed cases and deaths per region. Choropleth maps were produced using the tmap package, revealing substantial spatial variation in case distributions, with Greater Accra and Ashanti showing the highest concentration of infections. Visual inspection supported the need for spatial statistical modelling due to visible clustering patterns.

**Spatial Modelling Implementation:**

A spatial lag regression model (lagsarlm) was applied to examine the relationship between total COVID-19 cases and regional population counts. Neighbour relationships were established using queen contiguity, and a spatial weights matrix was constructed.

The model accounted for spatial autocorrelation in the dependent variable, with diagnostic outputs confirming significant spatial dependence. The residuals were mapped to assess model fit and identify regions with unexplained variance. The model assumption of spatial spill over was justified by the contiguous transmission nature of infectious diseases.

**Interpretation and Discussion:**

The spatial regression results indicated a strong positive relationship between regional population size and reported COVID-19 cases, validating that more populous regions bore a disproportionate burden of infection.

Spatial clustering observed in southern regions suggests urban density and mobility patterns contributed to higher transmission rates. Residual maps highlighted potential underreporting or unique regional factors influencing case counts in Northern and Volta regions. These insights underline the importance of resource allocation, targeted testing, and vaccination efforts in high-density urban centres.

**Conclusion:**

The spatial analysis effectively revealed regional disparities in COVID-19 incidence and demonstrated the utility of spatial econometric modelling in public health surveillance. Visual and statistical outputs support actionable recommendations for geographically targeted interventions. Findings were presented using interactive maps and diagnostic plots to enhance interpretability for stakeholders.