Comparing Time Series Forecasting Models for Health Indicators: A Clustering Analysis Approach

Cláudia Vinhal[1*], Alexandra Oliveira[1,2], Brígida Faria[1,2], Ana Paula Nascimento[1,3], Rui Pimenta[1,4]

1. ESS, Polytechnic of Porto, Portugal; 2. Artificial Intelligence and Computer Science Laboratory (LIACC member of LASI), Porto, Portugal; 3. Center for Translational Health and Medical Biotechnology Research (TBIO), Porto, Portugal; 4. Centre for Health Studies and Research of the University of Coimbra/Centre for Innovative Biomedicine and Biotechnology (CEISUC/CIBB), Coimbra, Portugal.

* Corresponding author email: 10220888@ess.ipp.pt

Introduction: Time series are the sequence of observations ordered by equal time intervals, crucial for understanding causality, trends, and forecasts. Its analysis can be applied to several areas, such as engineering, finance, and health (1,2). One problem with the time series study is clustering, mainly understanding when two parametric time series are considered similar (3). The sum of mortality and morbidity, referred to as "Burden of Disease", is measured by a metric called "Disability Adjusted Life Years" (DALYs) (4). These indicators are direct measures of health care needs, reflecting the global burden of disease in the population, and are crucial for public health study and surveillance (5). DALYs can be represented by Autoregressive Integrated Moving Averages (ARIMA) models, and in this context understanding clusters is crucial. **Objectives:** The primary goal is to compare different distance measures between ARIMA processes when used in clustering techniques. **Methods:** The study begins by exploring the temporal characteristics of DALYs, highlighting underlying patterns and trends. Then, ARIMA models are applied to represent and describe the time series. It's on this representation of the time series that the Piccolo, the Maharaj, and the LPC distance measures are applied to use clustering techniques and identify clusters. Additionally, 8 distinct cluster validation metrics are used. **Results:** Specific to 48 European countries, the results show that the choice of distance measure can greatly influence clustering outcomes and the number of clusters formed. While certain methods revealed geographic patterns, other factors, such as cultural or economic similarities, also influence cluster formation. These insights contribute to advancing the field of public health surveillance and intervention, ultimately aiming to alleviate the global burden of disease. **Conclusions:** This study offers insights into applying ARIMA processes in clustering techniques for analysing temporal health data. By comparing different distance measures, this research improves our understanding of underlying patterns and trends in health indicators over time.

Keywords: Distance Measures, Clustering, DALYs, ARIMA Models

References:

1. Palma W. Time Series Analysis. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2016.
2. Nielsen A. Practical Time Series Analysis. 1st Edition. Sebastopol, CA: O'Reilly Media Inc.; 2019.

3. Aghabozorgi S, Seyed Shirkhorshidi A, Ying Wah T. Time-series clustering – A decade review. Inf Syst. 2015 May 30;53:16–38.
4. Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2019 (GBD 2019). Seattle, United States: Institute for Health Metrics and Evaluation (IHME), 2020.
5. Mathers CD. History of global burden of disease assessment at the World Health Organization. Archives of Public Health. 2020 Aug 24;78(1).