Received: 12 February 2022 *Accepted: 28 April 2022*

Comparing Cumulative Trends (with an application to COVID-19 data)

Amaratunga, D. [1], Cabrera, J.[2], Ghosh, D.[2], Katehakis, M.[2], Wang, J.[2], Wang, W.[2], & Xyntarakis, M.[2]

damaratung@yahoo.com, cabrera@stat.rutgers.edu, debopriyag@gmail.com, mnk@rutgers.edu, jin.w@rutgers.edu, wenting.wang@business.rutgers.edu, mxyntarakis@business.rutgers.edu

[1]*Princeton Data Analytics, New Jersey, USA,*

[2]*Rutgers University, New Jersey, USA.*

Abstract

Many applications involve looking at and comparing trends in data. We will discuss some statistics that could be used to assess the similarity or dissimilarity between pairs of cumulative trends. These statistics can then be used to study sets of trends – for example, to cluster them or to compare them across different groups We will describe one possible approach and illustrate its use in a case study, in which we studied the trend over time of COVID-19 in New Jersey (NJ) in the USA. It was found that areas close to New York City had significantly different (more rapidly increasing) cumulative trends compared to areas further from New York City during the early days of the pandemic, but this difference dissipated as the pandemic progressed and spread within New Jersey itself. Overall, the method performed well and detected insightful differences. Various socio-economic factors could have influenced the spread of COVID-19 within NJ. It was also found that socio-economic factors which could have influenced the spread of COVID-19 within NJ are population, distance to NYC, and percent of low-income households. The dynamic nature of these relationships also needs to be studied, perhaps using extensions of the methodology discussed here.

Keywords: Covid-19, Hotelling's test, Manhattan distance, Multidimensional scaling.

Introduction

Almost all of data analysis is concerned with looking for and studying patterns in data. One such pattern that is often of interest is a trend. A trend can be thought of as follows. Given a set of univariate observations, $\{a_i\}$, a trend looks at how smoothly and systematically a_i shifts with increasing values of i. In particular, with time course data, i would be time and we would be interested in following how the $\{a_i\}$ pattern evolves over time.

In this paper we will discuss a method to study sets of trends. We shall focus on data where the $\{a_i\}$ are counts and we are interested in comparing and/or classifying cumulative trends. A cumulative trend $\{x_i\}$ is defined as follows. If a_{it} is the count at t for sample i, x_{it} is the sum of all counts up to t: $x_{it} = \sum_{k=1}^{t} a_{ik}$. We will describe several statistics that could be used to assess the dissimilarity between such monotonic trends and study their performance in a simulation.

Note that, while $\{a_i\}$ is arguably a time series, there are issues that would make it problematic to use conventional time series formulations

here to determine clusters or to classify them. In the applications of interest here, there are multiple series; all of them involve a fairly short time period and the data exhibits no seasonal or cyclical component. Also, even though the series of $\{a_i\}$ counts may be autoregressive, the correlation between successive elements may not remain fixed over time due to multiple reasons. Keeping in mind that it is ultimately the dissimilarity in the evolution of the counts that matters for our purpose here, conventional time series methodology will not be applied.

To motivate our approach, we shall give a case study involving COVID-19 data in which the procedure we describe has been successfully applied.

Materials and Methods

Given two cumulative trends, $x=\{x_i\}$ and $y=\{y_i\}$, suppose that we are interested in seeing how similar or dissimilar they are. There are several different statistics that could be used for this purpose as shown in Table 1 (Shirkhorshidi et al., 2015).

Table 1.
Different distances (statistics) used to measure similar / dissimilar between two trends.

Type of the distance	**Formula**
Euclidean distance	$d(x, y) = \sum_i (x_i - y_i)^2$
Manhattan distance	$d(x, y) = \sum_i \lvert x_i - y_i \rvert$
Maximum distance	$d(x, y) = \mathrm{Max}_i \lvert x_i - y_i \rvert$
Chi-squared distance	$d(x, y) = \sum_i (x_i - y_i)^2 / (x_i + y_i)$
Kullback-Leibler divergence	$d(x, y) = \sum_i (x_i - y_{ij}) \log(x_i, y_i)$

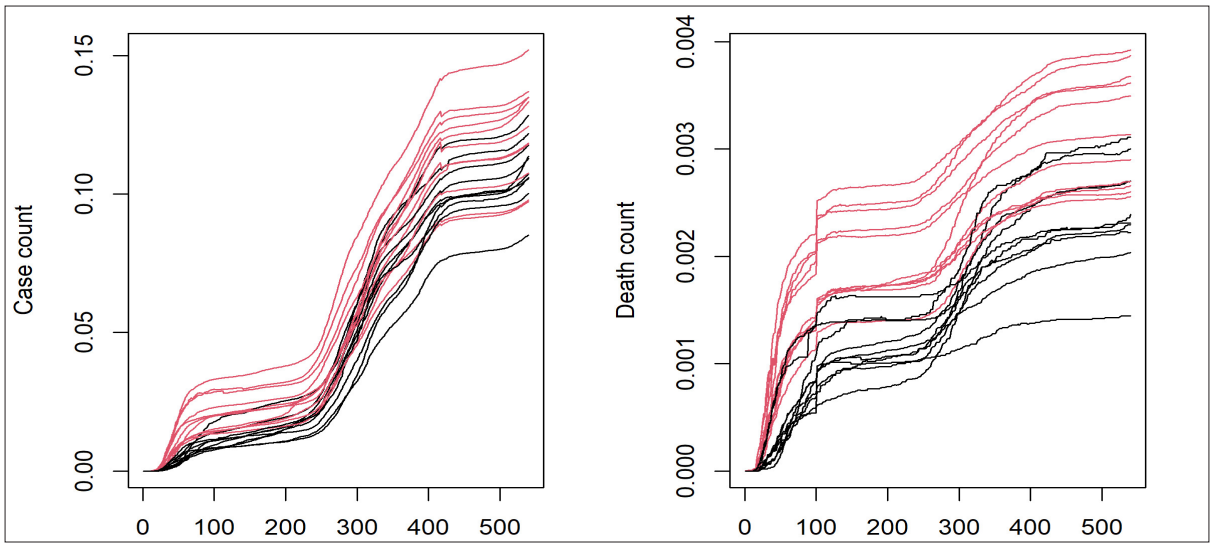
Once dissimilarities between pairs of trends have been calculated using one of these statistics, they can be analyzed in various ways as appropriate to the study objective (Johnson and Wichern, 2007). For instance, it is often helpful, when comparing multiple trends, to have a visualization that displays the similarity-dissimilarity patterns among the samples. This can be done using multidimensional scaling (MDS). Given *n* trends (x_i, $i=1, ...n$), MDS finds a set of n 2-dimensiional points $\{z_i\}$ which are such that the Euclidean distance z_i between z_j and approximates $d(x_i, x_j)$. Further analyses can be done using either the $\{z_i\}$ or the $\{d_{ij}\}$.

Case study: COVID-19 data

We studied the spread of COVID-19 across different geographic regions of New Jersey (NJ). NJ is a state located adjacent to New York City (NYC) in the USA where there is a lot of commuter and commercial traffic between NJ and NYC. We conjectured that NYC, being a highly populated metropolitan hub, would have a considerable influence on the spread of COVID-19 around NJ. To study this, we grouped the 21 counties of NJ into two groups: the 10 counties relatively close in commuting distance to NYC were designated as Group 1 and the 11 counties relatively distant from NYC were designated as Group2. In particular, we studied the cumulative time trends, $\{x_i\}$, of the counts of confirmed COVID-19 cases and COVID-19 related

deaths for each of the 21 counties of NJ for a period of approximately one and a half years, from 5 March 2020 to 26 August 2021 (Figures 1 and 2).

Figure 1.
Cumulative time trends, {x_i}, of the counts of confirmed cases (on the left) and deaths (on the right); each line corresponds to a county.



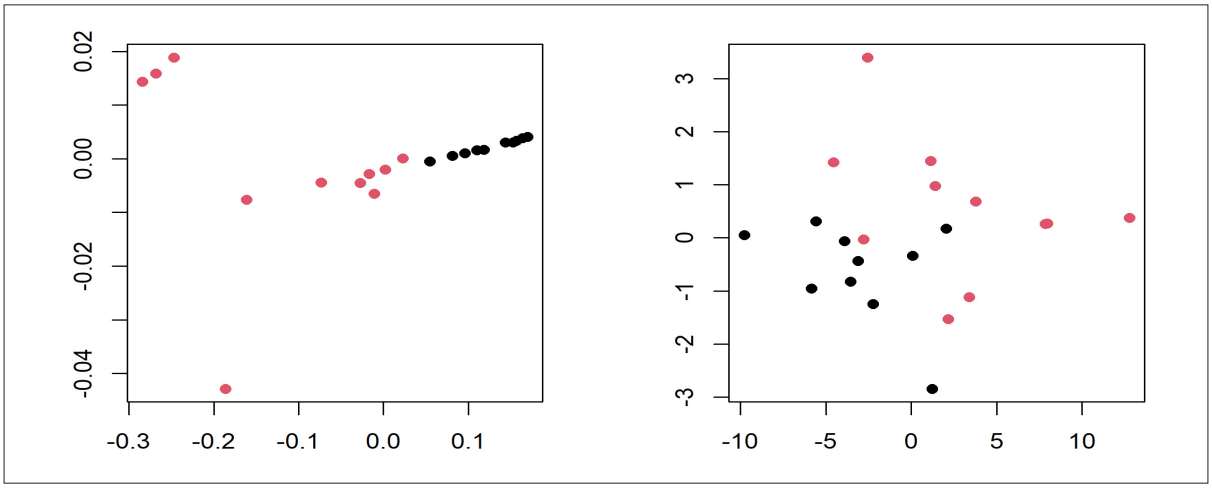
(Counties close to NYC are in red and counties distant from NYC are in black)

The COVID-19 pandemic evolved as a series of surges interspersed with periods of low counts for some months. Different regions had surges occurring at different times and with different characteristics. Since a surge has a pattern like a univariate distribution and since the simulation reported in Section 4 indicated Manhattan distance provided a good estimate of the separation between profiles that look like mixture distributions, we used Manhattan distance to estimate the dissimilarities between pairs of trends. Thus, we calculated the dissimilarities, D_{ij}, between all pairs of counties i and j.

$$D_{ij} = d(x_i/P_i, y_i/P_i),$$

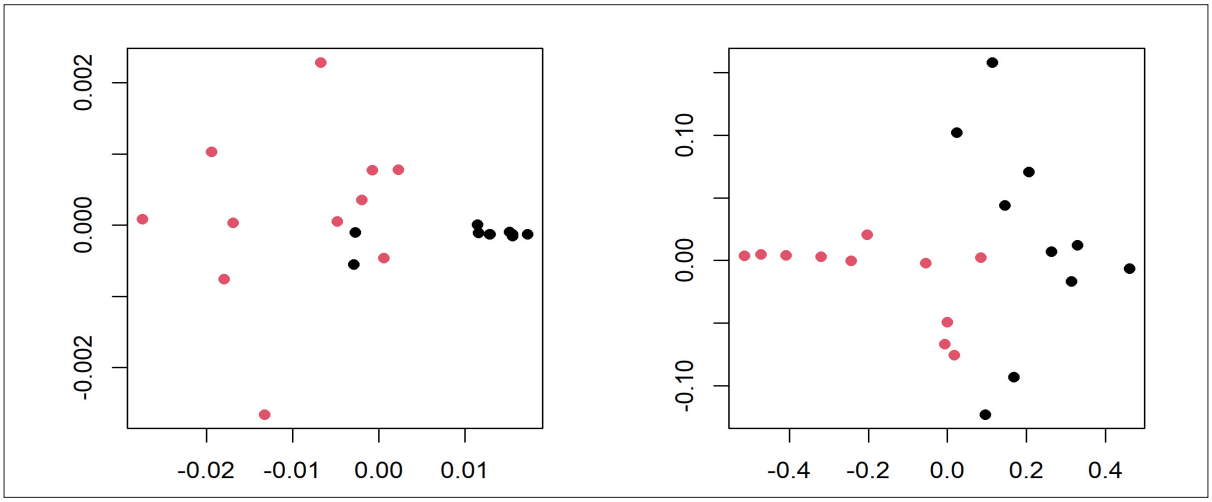
where P_i is the population of county i. This population adjustment was done to take into account the differences in county populations. Multidimensional scaling was performed using these dissimilarities. Figures 2 and 3 show MDS plots of the data (Figure 2 is for cases and Figure 3 is for deaths). We compared the counties relatively close in commuting distance to NYC (shown in red) to the counties relatively distant from NYC (shown in black). This was done for data up to Day 60 (plots on left) and for all the data (plots on right).

Figure 2.
MDS plots for the confirmed COVID-19 cases (The plot on the left is for the first 60 days and the plot on the right is for all 540 days).



(Counties close to NYC are in red and counties distant from NYC are in black)

Figure 3.
MDS plots for the COVID-19 deaths (The plot on the left is for the first 60 days and the plot on the right is for all 540 days).

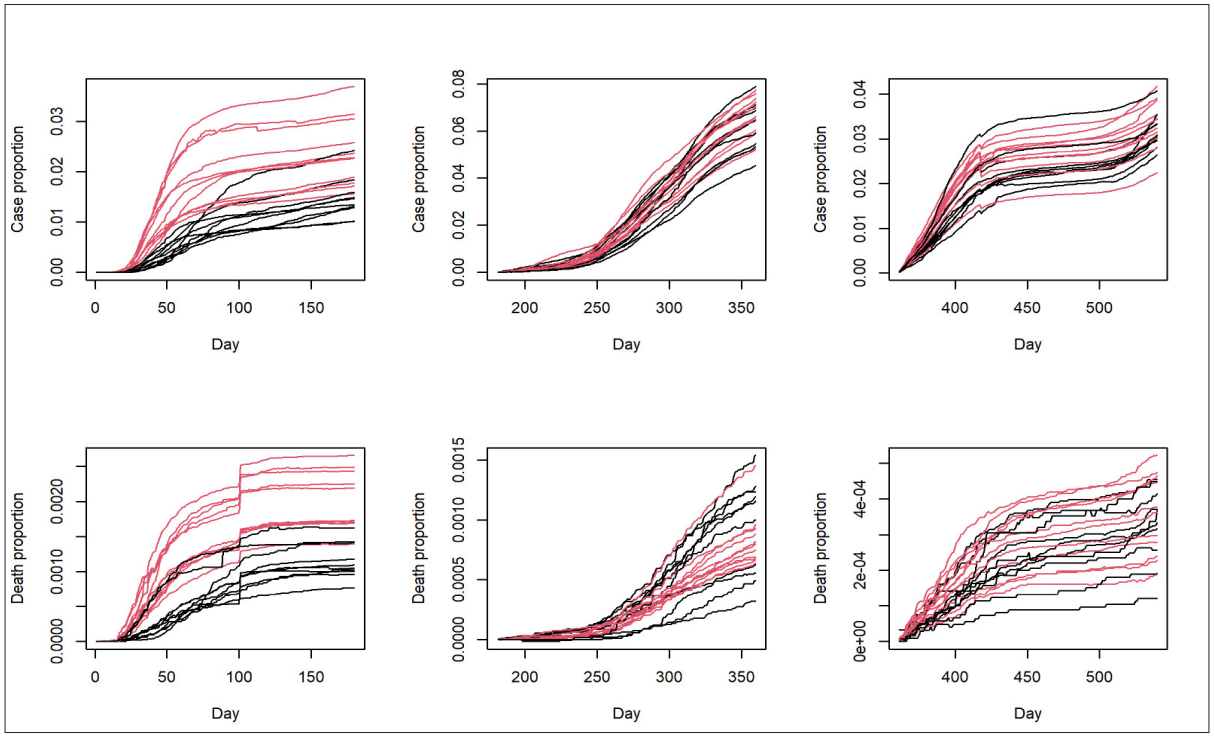


(Counties close to NYC are in red and counties distant from NYC are in black)

To study the time effect further, we separated the data into 3 six-month time periods: Days 1 to 180, Days 181 to 360 and Days 361 to 540 and looked at the cumulative time trends of the proportions of confirmed cases and deaths for each time period (Figure 4).

Figure 4.

Cumulative time trends, {x}, of the proportions of confirmed cases (top row) and deaths (bottom row) – from left to right, for days 1 to 180, days 181 to 360, and days 361 to 540. Each line corresponds to a separate county.



(Counties close to NYC are in red and counties distant from NYC are in black)

Table 2.

Hotelling's and t test p-values for cases and deaths.

Time period	For cases		For deaths	
	Hotelling p-value	T test p-value	Hotelling p-value	T test p-value
Days 1 to 180	0.0001*	0.0002*	0.0001*	0.0001*
Days 181 to 360	0.5339	0.7823	0.7849	0.8590
Days 361 to 540	0.0508	0.1854	0.1405	0.5830

We then carried out multidimensional scaling within each time period and applied Hotelling's test to compare the two groups of counties. We also took logs of the counts at the end of each time period and performed a t test. The results, shown in Table 2, indicate a very strong separation between the two groups of counties in the first six-month period, with significantly higher counts in the counties near NYC. However, this separation is no longer present in the later time periods, although there is perhaps a very slight shift again during the last six-month period. This demonstrates the geographic evolution of the pandemic over time, with many early cases and deaths in NJ arising from its proximity to NYC, but with this effect being dampened later as infections spread within NJ itself.

In related work, Amaratunga et al. (2021) also investigated whether various socio-economic factors could have influenced the spread of COVID-19 within NJ. They looked at factors such as percentage of elders in the population, percentage of low-income households, the numbers of food and health facilities (including fast-food and non-fast-food restaurants, groceries, nursing homes, fitness centers, and pharmacies – these were obtained by querying the Yelp Fusion API). Since a large proportion of the dissimilarities between counties could be accounted for by the first eigenvector of MDS (it was larger than 0.90 for both cases and deaths), it was reasonable to regard the values along this eigenvector as carrying the most information regarding differences between counties. These values were then used as a response variable and modeled against the above socio-economic factors. The model indicated that the important factors were population, distance to NYC, and percent of low-income households. The dynamic nature of these relationships is being studied.

Simulation

We carried out a simulation to compare the dissimilarity measures outlined in Table 1.

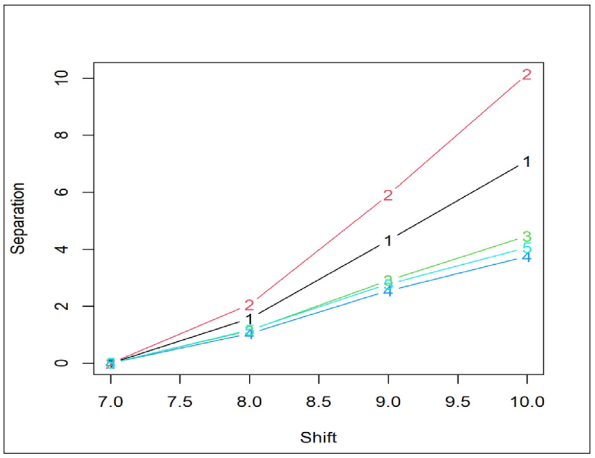
A set of 1000 observations was generated from a normal distribution with mean 7 and standard deviation 1. In addition, a set of 1000 observations was generated from a Mixed Normal distribution, in which 90% of the observations came from the same distribution as above and 10% of the observations came from a normal distribution with mean *m* and standard deviation 2. We then computed their empirical cumulative distribution functions and, based on these, calculated the dissimilarity d(m) between these two sets of observations using each of the different measures described in Section 2. This was repeated 500 times and the mean, $\overline{d(\mu)}$ was computed. This was then repeated for several different values of m.

Dissimilarity measures that are better able to discern the difference of the Contaminated Normal from the Normal will tend to have larger values of *d*. This can be used to compare the different measures. Since the dissimilarity measures are on different scales, the performance of a dissimilarity measure was evaluated using a standardized measure:

$$p(\mu) = \frac{\overline{d(\mu)} - \overline{d(0)}}{SD}$$

where *SD* is the standard deviation of the dissimilarities when m=0. Figure 5 shows a plot of separation, P(m) vsshift, m for the five dissimilarity measures.

Figure 5.
Performance measure P(m) vs m for the five dissimilarity measures. The lines are coded as follows: 1 = Euclidean distance, 2 = Manhattan distance, 3 = Maximum distance, 4 = Chi-squared distance, 5 = Kullback-Leibler divergence.



It can be seen that Manhattan distance most clearly captures the differences in cumulative trends as m increases.

Conclusions

We studied a method for comparing cumulative trends. A key aspect of the method is a dissimilarity measure. We compared the performance of five measures in a simulation and found that Manhattan distance displays the best performance in the setting in which the simulation was done.

A case study was also done and showed the value of the method. In the case study, we studied the trend over time of COVID-19 in New Jersey in the USA. It was found that areas close to New York city had significantly different (more rapidly increasing) cumulative trends in both confirmed cases and deaths compared to areas further from New York City during the early days of the pandemic, but this difference dissipated as the pandemic progressed and spread within New Jersey itself. Since this type of trend data arises in a variety of settings, this methodology is an overall useful tool to have.

References

Amaratunga, D., Cabrera, J., Ghosh, D., Katehakis, M. N., Wang, J., & Wang, W. (2021). Socio-economic impact on COVID-19 cases and deaths and its evolution in New Jersey. *Annals of Operations Research*, doi:10.1007/s10479-021-03941-4.

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*, 6th Edition, Pearson.

Shirkhorshidi, A. S., Aghabozorgi, S., & Wah, T. Y. (2015). A comparison study on similarity and dissimilarity measures in clustering continuous data, *PLoS ONE*, 10 (11), e0144059.