

COVID-19's daily new confirmed cases and deaths in Tunisia : A forensic analysis and machine learning algorithms for outlier detection

Ghassen El Montasser

Department of Quantitative Methods

ESC Business School of Tunis

University of Manouba

Campus Universitaire

2010 La Manouba

Tunisia

Abstract

Effective management of health crises requires a good diagnosis of the current situation and an update of relevant data. Therefore, reliability of this data is fundamental to implement the necessary measures to manage health crises. In this paper, we study the compliance of COVID-19 cases and deaths reported in Tunisia with Benford's law. Using a panoply of goodness-of-fit tests, we found that this compliance is disproved. This may raise doubts about the manipulation of this data. A search for outliers in these data followed this analysis using two machine learning algorithms, namely the local outlier factor (LOF) and isolation forests. The number of outliers in the data has heightened doubts about their reliability.

Subject Classification (2010): 62Q05, 68W99.

Keywords: COVID-19, Benford's law, Local outlier factor, Isolation forests.

1. Introduction

Since the World Health Organization (WHO) reported COVID-19 as a global pandemic, the symptoms of which were first observed in Wuhan, China, the world has seen millions of reported cases and a succession of waves attesting to the spread of the virus worldwide, some of which were more staggering than others. It will therefore be more than useful to touch

ORCID-ID: 0000-0003-1513-9630

E-mail: g.elmontasser@gmail.com

base with the statistics about the spread of the virus in the world. In this regard, over 200 territories reported two main statistics, namely, the number of new cases infected and the number of deaths caused by the virus. With the introduction of vaccines, the health situation around the world has improved, yet some countries have also postponed new daily vaccinations.

In recent years, numerous studies have concentrated on data about the COVID-19 pandemic. Along with various models employed in different papers on this topic, the research objectives vary. Certain studies have concentrated on predicting infection, mortality, recovery, and immunization trends against this outbreak. While the comparison of different forecasting models and the associated analysis of predictive uncertainty are rarely found in existing literature, some publications have delved into these key aspects, as the selection of the most appropriate models remains a subject of considerable debate within this analytical framework. Singh, Singh, and Khan, [1] analysed two key aspects and also presented real-time predictions of data about the COVID-19 outbreak through the utilisation of support vector machine models. An additional avenue of investigation within the analysis of the COVID-19 time series has been the assessment of the pandemic's economic consequences. It is widely acknowledged that the COVID-19 pandemic contributed to a slowdown in economic activity, particularly during lockdown periods. Consequently, numerous studies have investigated the effects of the pandemic on economic and financial variables, frequently employing time series models such as Time-varying VAR models [2,3]. Another body of research has delved into the social dimensions of this pandemic. In this context, Bhatnagar, Khan, and Singh, [4] conducted a detailed examination of COVID-19 data specific to India. They explored variables such as age, gender, travel history, mode of communication, and current status. The researchers identified a significant correlation between gender (male or female) and the mode of disease transmission.

Reliability of these figures is a key issue at the heart of any policy aimed at combating the coronavirus. However, public doubts about these figures fail them to be integrated into this type of policy in the absence of any societal conviction [5]. Data reliability analysis can be performed with Benford's law. The law is named after [6] introduced it although the outline of this law had already been stated [7]. However, Benford demonstrated in [6] what was enunciated by Newcomb in [7], and thereby enabled the great use of this law. Several researchers have conducted this forensic analysis on COVID-19 data using Benford's law. However, the results are

mixed. For example, Koliaş [8] found that most countries studied in his paper showed deviations from Benford's law. Such a finding may lead to doubts about the COVID-19 data. However, Farhadi [9] found that around 28% of countries well met Benford's predicted distribution, with only six countries having completely unreliable data. The diversification of the used goodness-of-fit tests and the samples considered may be behind these mixed results.

Hawkins [10] defined an outlier as «an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism». On the other hand, Aggarwal [11] stated that «in many cases such as intrusion or fraud detection, the outliers can only be discovered as a sequence of multiple data points, rather than as an individual data point. For example, a fraud event may often reflect the actions of an individual in a particular sequence ». Therefore, it is pellucid the close link that can subsist between the detection of anomalies in fraud analyses and that of outliers. The latter may indeed offer the analysis a richness and a deeper insight.

The contribution of this paper in the analysis of COVID-19 data anomalies is the dual use of outlier detection algorithms and Benford's law. More specifically, we will focus on two machine learning algorithms in outlier detection, namely the local outlier factor and isolation forests algorithms. In machine learning, these two algorithms have shown good properties in detecting outlier observations. To the best of our knowledge, no previous study has examined COVID-19 data using both forensic analysis and outlier detection models, through specialised algorithms. This joint use allows us to draw conclusions about the compliance of the data with Benford's law and the dates of the anomalies. Alternatively, examining these dates, we can determine what caused the data to deviate from its normal behaviour.

We will opt for forensic analysis and outlier detection to study COVID-19 new confirmed cases and deaths in Tunisia. We chose this country for several reasons. First, Tunisia has experienced, like all countries in the world, a succession of COVID-19 waves. During these waves, the country has had moments when it has managed the health crisis well. However, in other moments, the country has had lost control, sometimes dramatically. Second, the health infrastructure is in a kind of fragile state, which makes confronting COVID-19 a rather arduous task. Finally, Tunisia has experienced political instability, manifested by a succession of governments. As a result, it will be expected that there will be a lack of appropriateness in mapping out a health policy to improve the current health situation and combat exceptional health crises such as COVID-19.

The paper is organised as follows. In the second section, we will give a description of Benford's law and the two-outlier detection algorithms that we will use. The third section describes our findings obtained through the use of the forensic analysis and outlier detection algorithms in the Tunisian series of COVID-19 cases and deaths. The fourth section discusses the results. Finally, the fifth section concludes the paper.

2. Materials and Methods

2.1 Data

The data consist of COVID-19 cases and deaths in Tunisia and are obtained from the World Health Organization's website. More precisely, both series used in this paper, which are observed daily, represent new confirmed cases and deaths of this pandemic, recorded every day. The data cover the period from March 2, 2020, to August 12, 2022.

2.2 Conformity analysis with Benford's law

A probability distribution for identifying the first digit in a set of numbers is known as Benford's law (also known as the law of prime digits). After earlier research by the mathematician Simon Newcomb, KOLIAS [6] stated that in natural and unrestricted datasets, the likelihood of each digit appearing is given by the formula:

$$p(d) = \log_{10} \left(\frac{1+d}{d} \right), \quad \text{for } d = 1, 2, \dots, 9. \quad (1)$$

To test the null hypothesis of the conformity of the real data with Benford's law, i.e. when the data follow this law and then a certain natural distribution is checked, one can refer to goodness-of-fit tests. For example, if we choose the Chi-square test, we should use the following statistic to test the conformity null hypothesis:

$$X^2 = \sum_{i=1}^9 \frac{(O_i - E_i)^2}{E_i}, \quad (2)$$

where the digit is the index, O_i and E_i are the observed and expected frequencies for the i -th digit, respectively. However, it should be mentioned that the Chi-square test has some quandaries in large samples. In fact, in case of such samples, interpretation of the significance should be evaded, as this test has enough power to detect even diminutive deviations from the expected distribution [8,9].

A joint distribution of k significant digits of a positive real number m can be stated as follows

$$p(d_1, d_2, \dots, d_k) = \log_{10} \left(1 + \frac{1}{\sum_{h=1}^k 10^{K-h} d_h} \right), \quad (3)$$

where $d_1 \in \{1, 2, \dots, 9\}$ and $d_j \in \{0, 1, \dots, 9\}$ [12].

2.3 Outlier detection methods

2.3.1 Local outlier factor (LOF)

Breunig [13] have shown the sensitivity of the distance-based outlier detection methods to data locality. They introduced the Local Outlier Factor (LOF) approach to address this issue, which quantifies the outlierness of data points and allows for the adjustment of variances in various densities. More explicitly, density of data points and density of the data point neighbourhood are both considered by LOF. The LOF algorithm considers the relative density of the data point in its task of determining outliers. This function is the main feature of the algorithm. Data points will be sorted once the outlier scores have been calculated to identify outliers in the data set. The relative density of a data point x with k neighbours is given by the following equation:

$$\text{Relative density of } x = \frac{\text{Density of } x}{\text{Average density of all data points in the neighborhood}}. \quad (4)$$

The LOF algorithm goes through the following steps:

- 1) Calculate the distance between a point M and all given points using distances such as Euclidean or Manhattan.
- 2) Find the k -nearest neighbour closest point.
- 3) Find the k closest points.
- 4) Define the reachability distance (reachdist) between two points x and x' as

$$\text{reachdist}_k(x, x') = \max(\|x - x^{(k)}\|, \|x - x'\|), \quad (5)$$

where $x^{(k)}$ is the k th nearest neighbour of x in $\{x\}_{i=1}^N$ with N is the number of data points. Note that we have used the Euclidean distance in the above equation. Using this reachability distance, the algorithm will find the local reachability (lreachdist distance of x defined as:

$$\text{lreachdist}_k(x) = \left(\frac{1}{k} \sum_{i=1}^k \text{reachdist}_k(x^{(i)}, x) \right)^{-1}. \quad (6)$$

5) Finally the local outlier factor of x is defined as:

$$LOF_k(x) = \frac{\frac{1}{k} \sum_{i=1}^k \text{lreachdist}_k(x^{(i)})}{\text{lreachdist}_k(x)}. \quad (7)$$

The above steps, however, are subject to the following assumptions. The reachability distance is not symmetric. This can be understood by the fact that if x' is located in a dense region and if the distance between x and x' is large, the reachability distance between both points is equal to the true distance. On the other hand, if the distance between x and x' is small, their reachability distance will be smoothed by the k -nearest neighbour distance of x . The reachability distance can be seen as a stabilizer of the Euclidean distance $\|x - x'\|$ since the distance is not smaller by construction than $\|x - x^{(k)}\|$.

2.3.2 Isolation forests (iForest)

Liu, Ting, and Zhou [14] have introduced a method termed isolation forest, which shares some kindred perceptual facets with random forests, another set of techniques performing better than most classifiers across a range of domains. However, isolation forests differ, to a certain extent, in their assigning scores to distinct data points in an unsupervised way. A collection of isolation trees makes up an isolation forest. The key point of the isolation forest algorithm is the data's recursive and random partitioning. The aim is to isolate the instances in nodes containing fewer and fewer instances until we isolate the points in singular nodes containing only one instance. Because these outliers are in sparse locations, the branches of the tree including them will be less profound. An outlier's score is determined by the distance from the leaf to the root. iForest follows hence two stages: a training phase, which represents the creation of the forest, and a scoring phase. There are two control parameters in the training phase, and in the algorithm: (1) the sub-sampling rate ψ sets the number of samples used for each tree training, and (2) the number of trees nt , defining the complexity of the model. The next scoring phase starts by determining the outliers' average path length in their sparse regions. The path length, noted $h(x)$, of a point x is measured by the number of edges crossing a particular iTree from the root node until the process ends at an outer node. For a sample containing N points, the minimum height of the

tree is $\log(N)$. However, the maximum height is $N-1$. An anomaly score can be represented by s , that is:

$$s(x, N) = 2^{-\frac{E[h(x)]}{c(h(x))}}, \quad (8)$$

$$c(N) = 2H(N-1) - \left\lceil \frac{2(N-1)}{N} \right\rceil, \quad (9)$$

and

$$H(k) = \ln(k) + \varepsilon. \quad (10)$$

ε is an Euler constant, that is $\varepsilon = 0.5772156649$, while $E[h(x)]$ is the average of $h(x)$ in a collection of isolation trees. We have three cases:

- (a) if instances return s very close to 1, then they are definitely anomalies,
- (b) if instances have s much smaller than 0.5, then they are quite safe to be regarded as normal instances, and
- (c) if all the instances return $s \approx 0.5$, then the entire sample has no distinct anomalies.

There is a close link between subspace outlier detection and isolation forests. To clarify, the branches represent local regions of the data subspaces depending on how attributes are chosen for splitting. The outliers will thus be found in the lower dimensional subspaces with shorter paths [11].

When it comes to real data with a high number of observations, iForest compared to LOF can provide gains in execution time. This is because the complexity of the learning phase is given by $O(nt\psi \log \psi)$ in iForest algorithms, while by contrast LOF algorithms have quadratic complexity (i.e. $O(N^2)$). The complexity of iForest algorithms is defined in this way since each element of the ψ data elements of the nt trees should be isolated or quasi-isolated in the associated tree [15].

3. Results

3.1 Data Description

The graphical representations of the new confirmed cases and deaths show an alternation between steep fluctuations and smooth fluctuations.

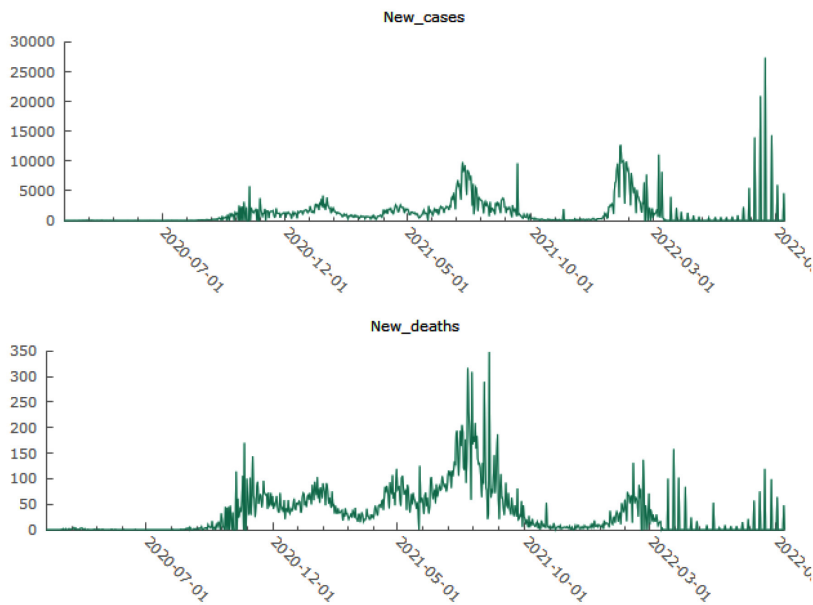


Figure 1
Graphics of COVID-19 data

Figure 1 shows the two graphs of these two series. The peaks of COVID-19 cases are observed in July 2022. However, death peaks are observed around July and August 2021. This trend resumed around March 2022.

Table 1
Summary statistics

Series	Mean	Median	Standard Error	Skewness	Kutosis	χ^2 -test for one outlier	Grubbs test for one outlier
New cases	1273.744	426	2288.879	4.3062	32.7317	129.69***	11.3882***
New deaths	32.5917	11	45.3955	2.3044	11.045	48.275***	6.94800***

Notes: *** denotes rejection at the 1% level of the null hypothesis that the largest value is not an outlier.

Table 1 gives the summary statistics for the two series of COVID-19 cases and deaths. The two series are normally distributed since they show positive skewness and their Kurtosis statistics are considerably higher than three. We run the Chi-square and Grubbs tests to conclude whether the larger values of the two series are outliers. To know about the two tests [16,17]. The results show clearly that the largest values of both series are outliers.

3.2 Benford's law

To determine whether the two series conform to Benford's law, we used the Chi-square test. Nevertheless, since this test has some limitations, as mentioned above, we also used Kuiper, Mantissa Arc and mean absolute deviation (MAD) tests [18]. Koch and Okamura [19] argued for the Kuiper test as a component of the Benford's Law compliance analysis. However, we should mention at this point that this test is invariant to cyclic transformations in the series. Consequently, the test is highly unreliable in case of variations by season of the year, day of the week or intra-daily, etc. There is no convention on the preponderance of a goodness-of-fit test associated with a Benford analysis. Therefore, we opted for a variety of tests to further investigate the reliability of the studied series. To this end, we used the "benford.analysis" R package. In addition, we performed all these tests on the entire sample, lockdown and post-lockdown samples of both series. Some clarifications here are in order. We have chosen the total lockdown of 2020, which has taken place in Tunisia for quite a long period. After this lockdown, some micro-lockdowns took a few days. Their consideration in this conformity analysis is not correct since they do not show the required range of data. The total lockdown considered in this analysis lasted from 18/03/2020 to 07/03/2020. Although the Tunisian President officially announced the lockdown on 20/03/2020 in a televised speech, we chose the start date of the lockdown on 03/18/2020, where a curfew was announced, courses stopped and the president of the government started talking about the intention to impose a lockdown. However, the date of 03/07/2020 represents the release of a government decree officially announcing the end of the total lockdown. It is worth noticing that the number of days of the pre-lockdown period is not very large and does not provide the ideal framework for the application of the Benford's law. Unlike Koch and Okamura [19], we considered the in-lockdown period, which in the Tunisian context can be informative in analysing the conformity of COVID-19 data with the Benford's law for

Table 2
First digit distribution and tests of significance

		Leading digits									N	X ² -Stat	Mantissa Arc Test	Kuiper	MAD	MAD Conformity	Distorsion Factor
		1	2	3	4	5	6	7	8	9							
New cases	Full Sample	281	136	66	42	50	39	38	32	29	713	41.555***	0.0307***	2.693734**	0.0239	Nonconformity	-12.7870
New cases	In-Lockdown	29	19	13	9	8	4	1	2	2	87	7.9705	0.0206	2.960503***	0.0263	Nonconformity	-35.1371
New cases	Post-Lockdown	250	115	53	31	40	35	37	30	26	617	46.862***	0.0390***	2.861747**	0.0259	Nonconformity	-11.3391
New deaths	Full Sample	140	75	81	72	65	62	50	36	23	604	48.059***	0.0459***	3.22151***	0.0286	Nonconformity	10.0877
New deaths	In-Lockdown	11	7	5	3	3	2	0	1	1	33	26.11***	0.2458***	17,62035***	0.0825	Nonconformity	-
New deaths	Post-Lockdown	118	69	77	71	65	62	50	36	23	571	65.686***	0.0545**	3.878216***	0.0345	Nonconformity	10.0877

Notes: ^{***} denotes rejection, at the 1% level, of the null hypothesis that the observed distribution follows the theoretical (Benford) distribution.

several reasons. First, for approximately one month during the in-lockdown period, Tunisian health authority used a single laboratory to test for COVID-19 infections in the entire country. Therefore, it is appropriate to study the conformity of COVID-19 cases under these circumstances. The health authorities chose the decentralisation of the laboratories subsequently. Second, in-lockdown samples show a large number of observations (about 471 observations) and a very rich range with pronounced variations, especially for COVID-19 cases.

Table 2 shows that with all four tests, the null hypothesis of conformity with the Benford’s law is rejected for both studied series and their sequences except for COVID-19 cases during the in-lockdown period, where the null hypothesis is accepted with the Chi-square and Mantissa Arc tests. Our results are inconsistent with [9]. This may be due, as mentioned in the introduction, to the goodness-of-fit tests used and the samples considered.

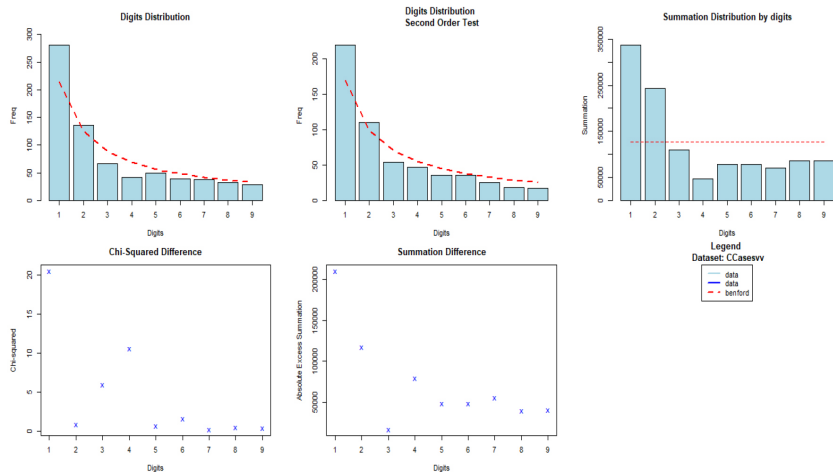


Figure 2

The distribution of the first digits of the COVID-19 cases (full sample)

Figures 2 and 3 show the distributions of the first and second digits for COVID-19 cases and deaths, respectively. We notice that in the case of the first and second digits, there is a clear divergence from the expected Benford’s law digits.

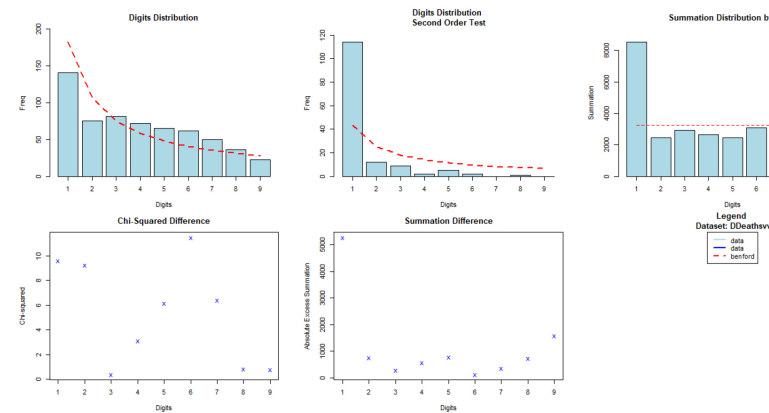


Figure 3
The distribution of the first and second digits of the COVID-19 deaths (full sample)

3.3 Outlier detection

As mentioned above, Table 1 shows that the largest values of both series are two outliers when referring to the Chi-square and Grubbs tests. Taking this descriptive analysis further, we have plotted the box plots of the two series. As shown in Figure 4, the box plots highlight outliers in both time series. We have also considered a normal bivariate distribution

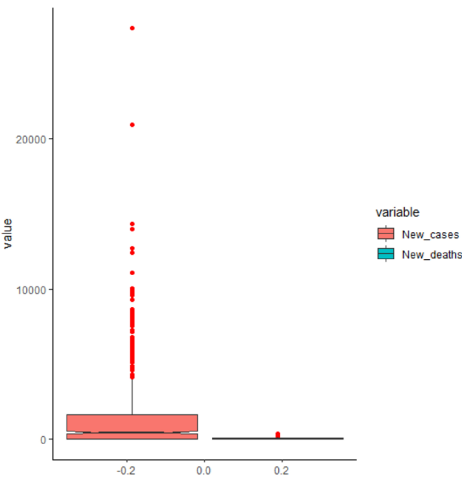


Figure 4
Boxplots of COVID-19 cases and deaths

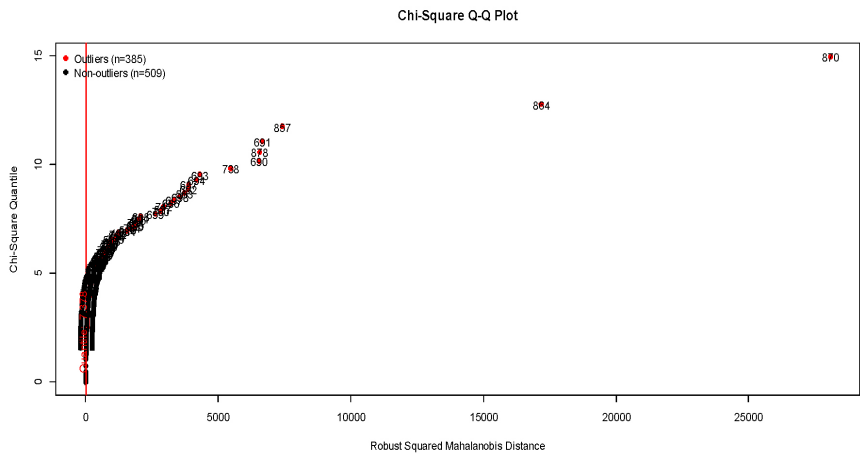


Figure 5
Chi-square Q-Q plot of COVID-19 cases and deaths

of the vector of both series. Figure 5 shows the Q-Q plot of this distribution, considering the robust squared Mahalanobis distance. Again, the outliers are remarkably observed where their observation numbers are also known.

From the descriptive analysis, we can deduce that there are outliers in both series. Before applying the LOF algorithm, we eliminated duplicate observations since the algorithm may have problems with their presence.

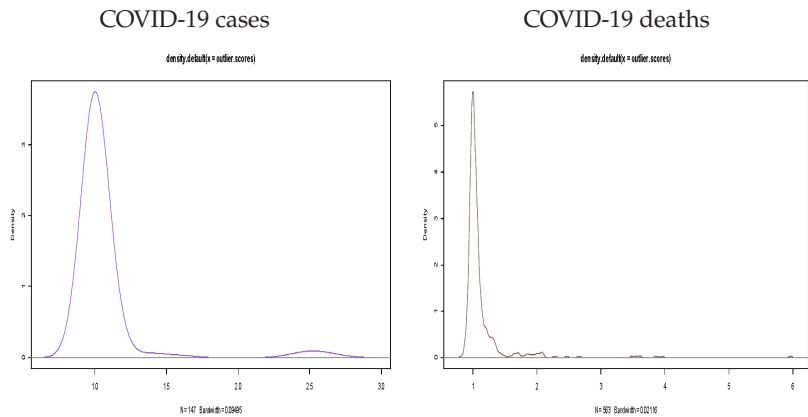


Figure 6
Outliers scores of COVID-19 cases and deaths

To run this algorithm, we used the R package “Rlof”. Figure 6 shows the different scores of the different points of the two series. We have chosen the points whose scores exceed those of the 97% order quantiles as outliers.

Table 3 gives the dates of the outliers for both series. We notice that COVID-19 cases exhibit a higher number of outliers than COVID-19 deaths. About 76.5% of the outliers in the first series are observed from 11/07/2021 onwards. However, death series has only five outliers, three of which are observed in July and two in August 2021. To detect outliers with isolation forests in the studied series, we used the “isotree” R package.

Table 3
Dates of outliers from local outlier factor (LOF) algorithm

New cases	New deaths
14-12-2020	24/07/2021
28-12-2020	25/07/2021
03-02-2021	29/07/2021
13-02-2021	13/08/2021
11-07-2021	19/08/2021
12-08-2021	-
18-08-2021	-
23-09-2021	-
26-09-2021	-
19-01-2022	-
20-01-2022	-
15-02-2022	-
08/03/2022	-
06/07/2022	-
13/07/2022	-
19/07/2022	-
27/07/2022	-

Albeit the isolation forest algorithm may have some failures in the detection of local outliers, it has strengths that explain its preponderance perceived by a great number of researchers in data mining. Isolation forests have indeed advantages over the LOF algorithm in terms of some aspects cognate to the insistence and robustness to outliers in the training

data, multimodal distributions, and scale changes of the variables. Additionally, isolations are faster to fit because of the LOF algorithm's above-mentioned complexity, which results in a long output delivery time [20]. The averages of the various lengths between the root node and the outside nodes for both series using isolation forests are displayed in Figure 7. These plots provide a preliminary understanding of which types of points are more or less prevalent than the others.

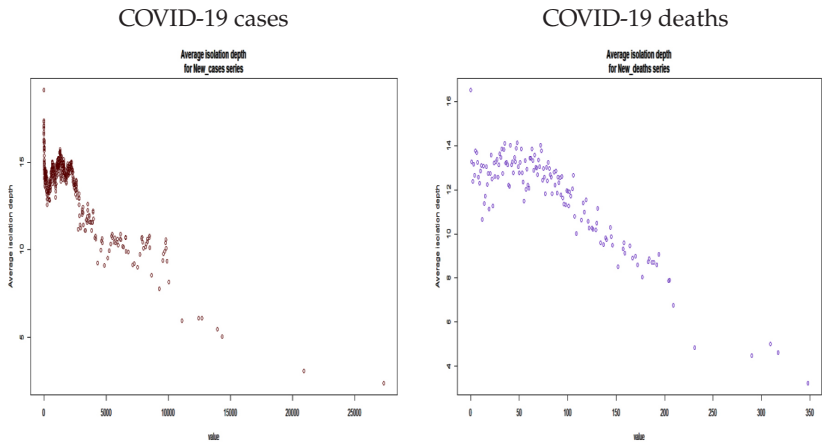


Figure 7

Average isolation depth of COVID-19 cases and deaths

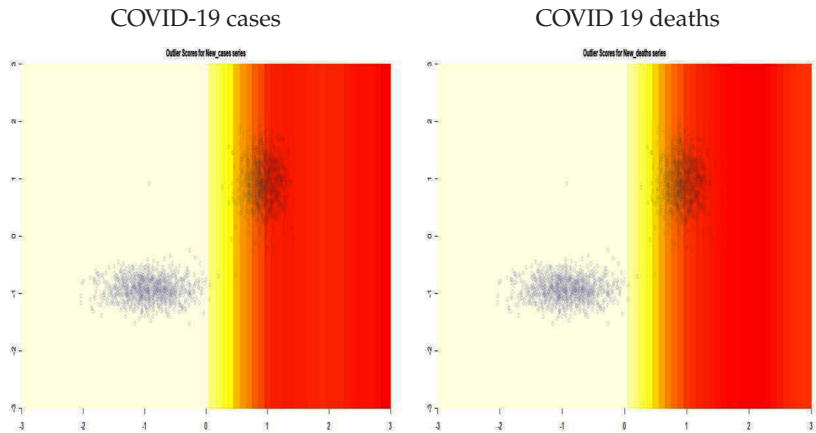


Figure 8

Outlier scores for COVID-19 cases and deaths

Table 4
Dates of outliers from isolation forest algorithm

New cases	New deaths
11/07/2021	27/10/2020
13/07/2021	10/07/2021
18/07/2021	11/07/2021
12/01/2022	14/07/2021
16/01/2022	15/07/2021
17/01/2022	17/07/2021
19/01/2022	18/07/2021
20/01/2022	21/07/2021
22/01/2022	24/07/2021
23/01/2022	25/07/2021
24/01/2022	26/07/2021
28/01/2022	29/07/2021
29/01/2022	30/07/2021
08/03/2022	31/07/2021
06/07/2022	01/08/2021
13/07/2022	02/08/2021
19/07/2022	03/08/2021
27/07/2022	04/08/2021
-	13/08/2021
-	19/08/2021
-	28/08/2021
-	29/08/2021

Figure 8 shows the different scores obtained by forest isolations. Referring to equations 8, 9, and 10 and fixing a preliminary threshold, one can optically discern which points are outlier observations. In this regard, we have chosen the identification threshold of 0.6, following [14]. Then, points with scores higher than 0.6 are outliers. Table 4 shows the dates of the outliers in both series using the isolation forest. We notice that the latter algorithm revealed a larger number of outlier observations than the LOF algorithm, especially for the death series. The isolation forest algorithm has detected eighteen outliers in COVID-19 cases against seventeen detected by the LOF algorithm for the same series. However, the

isolation forest algorithm found twenty-two outliers in COVID-19 deaths against five outliers detected by LOF for the same series. The LOF and forest isolation algorithms share eight dates representing outliers in COVID-19 cases. These dates are as follows: 11/07/2021, 19/01/2022, 20/01/2022, 08/03/2022, 06/07/2022, 13/07/2022, 19/07/2022, and 27/07/2022. It will be worth noticing that the isolation forest algorithm also detected the dates of the outliers detected already by the LOF algorithm in the death series. However, as mentioned above, the latter algorithm found a higher number of outliers, and hence dates, in this series. Further scrutiny is required to deeply analyse the results on the death series. This should be among the subject of the following subsection.

3.4 Discussion of the results

As pointed out by [5], non-compliance of the COVID-19 series with the Benford's distribution does not obligatorily denote a choice to defraud data. This may be because of national public health policies that leave much to be desired, of poor capacity management or reporting. We should recall that the Tunisian public health authorities used only one laboratory to detect suspected individuals for approximately the first month of the pandemic. COVID-19 has had several variants where each time the rate of infections spread will be more stressed. Among the latest variants is OMICRON, which is associated with a very high transmission rate, but not causing serious symptoms and not requiring long recovery periods like the previous variants. As a result, a sizable proportion of the population does not report positive expeditious antigen tests to the health authorities.

However, the political changes, manifested especially by the recurrent successions of governments that Tunisia has experienced in the last ten years, may provide further insights into our results. This has led to political instability, which has created a kind of economic vulnerability. Under such circumstances, the decision to fully lock down certainly carried enormous economic and social costs. First, to implement this decision, the country has suffered an unprecedented supply shock. Second, hundreds of thousands of citizens, working in informal and precarious jobs, have been deprived of any source of income. The authorities, therefore, to avoid public anger, might have moved towards a tendency to increase the number of deaths to tell the people that their health was in danger and the measures taken were worthwhile. This may explain the significant

evidence of non-compliance of COVID-19 deaths during the in-lockdown period. On the other hand, the country wanted to send a message to the outside world that, after a reasonable time span, it has controlled a particular wave of coronavirus. The challenge here is to attract foreign direct investments and encourage tourist arrivals. Machine learning algorithms accurately provided the dates of outliers in both series. The main message of this study is that changes in government or political system did not help ease the previous situation.

To summarize, if there are doubts about the reliability of the data, this may be due to two reasons: the first reports to the frangibility of the health infrastructure and the specificity of individual data. However, the second relates to some political-socio-economic requisites or factors.

4. Conclusion

Deviations from the Benford's distribution do not necessarily insinuate that the data is fraudulent. However, the analysis of compliance with this distribution can provide a preliminary insight into the possibility of data manipulation. In this paper, we studied outliers in Tunisian COVID-19 cases and deaths using two machine learning algorithms, namely LOF and isolation forests. We found a number, which is not minute, of outlier observations that may fuel doubt about the reliability of these data. The dates of the outliers show that Covid-19 data reporting is not only a function of a change of a government or a political system, but it depends on the health infrastructure and operational continuity under the framework of well-determined objectives and long-term health policy.

References

- [1] V. Singh, R. P. Singh, and M. A. Khan, "Prediction of COVID-19 corona virus pandemic based on time series data using support vector machine," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 23, no. 8, pp. 1583–1597 (Nov. 2020), doi: <https://doi.org/10.1080/09720529.2020.1784535>.
- [2] Z. Umar, S. Aziz, and D. Tawil, "The impact of COVID-19 induced panic on the return and volatility of precious metals," *Journal of Behavioral and Experimental Finance*, vol. 31, p. 100525 (Sep. 2021), doi: <https://doi.org/10.1016/j.jbef.2021.100525>.

- [3] G. E. Montasser, R. Gupta, J. Charl, and S. M. Miller, "The Time-series Linkages between US Fiscal Policy and Asset Prices," *Public Finance Review*, vol. 48, no. 3, pp. 303–339 (Apr. 2020), doi: <https://doi.org/10.1177/1091142120916032>.
- [4] V. Bhatnagar, M. A. Khan, and R. P. Singh, "Descriptive analysis of COVID-19 patients in the context of India," *Journal of Interdisciplinary Mathematics*, vol. 24, no. 3, pp. 489–504 (Jun. 2020), doi: <https://doi.org/10.1080/09720502.2020.1761635>.
- [5] N. Farhadi and H. Lahooti, "Forensic Analysis of COVID-19 Data from 198 Countries Two Years after the Pandemic Outbreak," *COVID*, vol. 2, no. 4, pp. 472–484 (Mar. 2022), doi: <https://doi.org/10.3390/covid2040034>.
- [6] F. B. Benford, "The law of anomalous numbers," *Proceedings of the American philosophical society*, pp. 551–572 (1938).
- [7] S. Newcomb, "Note on the frequency of use of the different digits in natural numbers," *American Journal of mathematics*, vol. 4, no. 1, pp. 39–40 (1881).
- [8] P. Kolias, "Applying Benford's law to COVID-19 data: the case of the European Union," *Journal of Public Health*, vol. 44, no. 2, pp. e221–e226 (Mar. 2022), doi: <https://doi.org/10.1093/pubmed/fdac005>.
- [9] N. Farhadi, "Can we rely on COVID-19 data? An assessment of data from over 200 countries worldwide," *Science Progress*, vol. 104, no. 2, p. 003685042110212 (Apr. 2021), doi: <https://doi.org/10.1177/00368504211021232>.
- [10] D. M. Hawkins, *Identification of Outliers*. Dordrecht: Springer Netherlands (1980). doi: <https://doi.org/10.1007/978-94-015-3994-4>.
- [11] C. C. Aggarwal, *Outlier Analysis*. Cham: Springer International Publishing (2017). doi: <https://doi.org/10.1007/978-3-319-47578-3>.
- [12] A. Cerasa, "Testing for Benford's Law in very small samples: Simulation study and a new test proposal," *PLOS ONE*, vol. 17, no. 7, p. e0271969 (Jul. 2022), doi: <https://doi.org/10.1371/journal.pone.0271969>.
- [13] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," *ACM SIGMOD Record*, vol. 29, no. 2, pp. 93–104 (Jun. 2000), doi: <https://doi.org/10.1145/335191.335388>.

- [14] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," 2008 Eighth IEEE International Conference on Data Mining (Dec. 2008), doi: <https://doi.org/10.1109/icdm.2008.17>.
- [15] Y. Chabchoub, M. U. Togbe, A. Boly, and R. Chiky, "An In-Depth Study and Improvement of Isolation Forest," *IEEE Access*, vol. 10, pp. 10219–10237 (2022), doi: <https://doi.org/10.1109/ACCESS.2022.3144425>.
- [16] W. J. Dixon, "Analysis of Extreme Values," *The Annals of Mathematical Statistics*, vol. 21, no. 4, pp. 488–506 (Dec. 1950), doi: <https://doi.org/10.1214/aoms/1177729747>.
- [17] F. E. Grubbs, "Sample Criteria for Testing Outlying Observations," *The Annals of Mathematical Statistics*, vol. 21, no. 1, pp. 27–58 (Mar. 1950), doi: <https://doi.org/10.1214/aoms/1177729885>.
- [18] M. Morzy, T. Kajdanowicz, and B. K. Szymański, "Benford's Distribution in Complex Networks," *Scientific Reports*, vol. 6, no. 1 (Oct. 2016), doi: <https://doi.org/10.1038/srep34917>.
- [19] C. Koch and K. Okamura, "Benford's Law and COVID-19 reporting," *Economics Letters*, vol. 196, p. 109573 (Nov. 2020), doi: <https://doi.org/10.1016/j.econlet.2020.109573>.
- [20] Z. Cheng, C. Zou, and J. Dong, "Outlier detection using isolation forest and local outlier factor," *Proceedings of the Conference on Research in Adaptive and Convergent Systems* (Sep. 2019), doi: <https://doi.org/10.1145/3338840.3355641>.

Received January, 2024