PLOS DIGITAL HEALTH

RESEARCH ARTICLE

Uncertain imputation for time-series forecasting: Application to COVID-19 daily mortality prediction

Rayane Elimam, Nicolas Sutton-Charani*, Stéphane Perrey, Jacky Montmain

EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Ales, Ales, France

These authors contributed equally to this work.
* nicolas.sutton-charani@mines-ales.fr

Check for updates

OPEN ACCESS

Citation: Elimam R, Sutton-Charani N, Perrey S, Montmain J (2022) Uncertain imputation for time-series forecasting: Application to COVID-19 daily mortality prediction. PLOS Digit Health 1(10): e0000115. https://doi.org/10.1371/journal.pdig.0000115

Editor: Rutwik Shah, UCSF: University of California San Francisco, UNITED STATES

Received: April 27, 2022

Accepted: August 30, 2022

Published: October 25, 2022

Copyright: © 2022 Elimam et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The data used in this article can be accessed at the public website https://ourworldindata.org/coronavirus.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have no competing interests to declare.

Abstract

The object of this study is to put forward uncertainty modeling associated with missing time series data imputation in a predictive context. We propose three imputation methods associated with uncertainty modeling. These methods are evaluated on a COVID-19 dataset out of which some values have been randomly removed. The dataset contains the numbers of daily COVID-19 confirmed diagnoses ("new cases") and daily deaths ("new deaths") recorded since the start of the pandemic up to July 2021. The considered task is to predict the number of new deaths 7 days in advance. The more values are missing, the higher the imputation impact is on the predictive performances. The Evidential *K*-Nearest Neighbors (E*K*NN) algorithm is used for its ability to take into account labels uncertainty. Experiments are provided to measure the benefits of the label uncertainty models. Results show the positive impact of uncertainty models on imputation performances, especially in a noisy context where the number of missing values is high.

Author Summary

The methodological aim of this study was to take advantage of missing data chronology in the imputation process in order to handle missing time series data. The practical goal of COVID application was to study the link between the numbers of chronological COVID confirmed cases and death. To achieve these goals we proposed 3 imputation methods of missing time series data each of them associated with an uncertainty model. For the COVID number of death prediction task, we set up a non-linear regression modeling for the number of COVID deaths prediction from past deaths and confirmed cases data. This led us to extend the Evidential K-Nearest Neighbor method to regression problems and to assess the impact of uncertainty modeling within imputation process in regards to predictive task. Finally, we showed the superiority of the time-E*K*NN (TEKNN) in terms of predictive performances compared to the Last Observation Carried Forward (LOCF) and Centered Moving Average (CMA) methods. More globally, we showed the interest of

modeling the uncertainty in the imputation process to better control the prediction error, especially during relative stable periods.

1 Introduction

With an increasing number of machine learning applications, data availability is becoming very important. Yet available datasets are often incomplete due to different measurement failures, especially when the data collection involves human participation. The treatment of missing values for predictive tasks has become an important issue giving rise to a wide range of research. Many methods have been proposed to handle missing values (average, omission, learning, etc.), one of the most popular being simply to exclude incomplete examples from the learning set, due to the incapacity to deal with missing values of most predictive models [1, 2]. That type of treatment remains undesirable with limited amounts of available data or in a chronological data structure.

The chosen method also depends on the nature of the missing values, which is often categorized in *Missing At Random* (MAR) for missing values that are dependent on observed values, *Not Missing At Random* (NMAR) missing values which depend on unobserved values and *Missing Completely At Random* (MCAR) missing values which are independent of observed or unobserved values [3, 4]. Those categories indicate why data are missing, an information to be taken into account in the imputation method [5].

Moreover, in a time-series forecasting context, missing values introduce irregular time stamps that contradict the most common hypothesis of standard time series methods. In terms of uncertainty, missing values can be interpreted as total ignorance or complete imprecision about the actual values. Some soft computing methods are designed to handle data uncertainty by modeling its degree [6–8]. In such frameworks, ignorance corresponds to the highest level of uncertainty and therefore missing values can be incorporated in models that take into account the uncertainty level. In this study our aim is to predict COVID-19 daily deaths in an artificially noised dataset out of which some labels (number of new deaths) have been randomly removed, resulting in MCAR missing values since the missingness is not related to any observed or unobserved values. The benefits of associating uncertainty models to imputation methods are studied. We evaluate the predictive performance of the Evidential-*K* Nearest Neighbors algorithm once missing data are imputed with and without uncertainty models (in the latter case the imputed labels are considered as certain).

The structure of the dataset is adapted to time series forecasting. We propose a methodology to handle the uncertainty inherent to missing values imputation methods. Some theories allow representation of uncertainty in a broader way than classical probability theory. Missing values uncertainty can be handled in different frameworks, *e.g.* fuzzy sets [9], possibilities distribution [10], probability sets [11], belief functions [12, 13]. We chose the belief functions framework for its flexibility and relative simplicity and also because recognized machine learning algorithms based on that framework are available [14–17].

Beyond standard machine learning researches on missing data imputation methods [1, 2], some soft computing imputation methods have been proposed [18–20]. In [21], a method is proposed to categorize missing data and to remove noise with a kernel-based approach that enables classification within the belief function framework. The purpose of the method is to design an imputation strategy providing uncertainty *resistance*; however the method does not handle the uncertainty at the predictive level. In [22] the authors propose a method to minimize the classification errors due to uncertainty caused by missing values. Multiple precise

missing values estimations are performed and the corresponding predictions are finally combined in predictive belief functions. In the context of information retrieval, Jousselme *et al.* proposed a missing values uncertainty representation [23]. Missing data are modeled as a belief function defined over the variables spaces. The method shows good performance for information retrieval task. As a matter of fact, none of those methods allows for the taking into account the uncertainty associated with imputation at the predictive level. In this study, we propose an approach to impute missing data in a chronological dataset and to model the resulting uncertainty in the belief functions framework. Finally an evidential classification model (E*K*NN) is extended to regression tasks in order to take into account the uncertainty associated with the imputation process.

The rest of this paper is organized as follows: first we present the main results of this study in Section 2, then we present our conclusion and perspectives in Section 3. All the details of our approach are given in Section 4 where we briefly recall the basis of the belief functions framework basis and the E*K*NN algorithm in the first subsection 4.1. After the presentation of the time series forecasting problem in an incomplete data context in the following subsection 4.2, three missing value imputation methods are proposed in subsection 4.3. In subsection 4.4 we present the uncertainty models associated with the previously introduced imputation methods; The uncertainty we are handling in this study is epistemic as we have no information about the missing label values. The chosen predictive model is Evidential-*K* Nearest Neighbors for its simplicity and its ability to deal with uncertain labels [14].

2 Results

First, we observe in Fig 1 that the three imputation methods are comparable in terms of performance. The TEKNN approach seems to perform better than the LOCF and CMA methods and its superiority grows as the noise level increases. Except for a small noise level of 0.1, the TEKNN model seems to be the best imputation method.

On the chronological evaluation with a noise level of 0 (Fig 2), we observe that the E*K*NN predictions with and without uncertainty models associated to label imputation blend together. This observation was expected as the data are not noised, *i.e.* there is no uncertainty associated with training labels. During the increasing and decreasing phases of the number of deaths, the E*K*NN seems to perform better than the baseline approach (blue and green curves

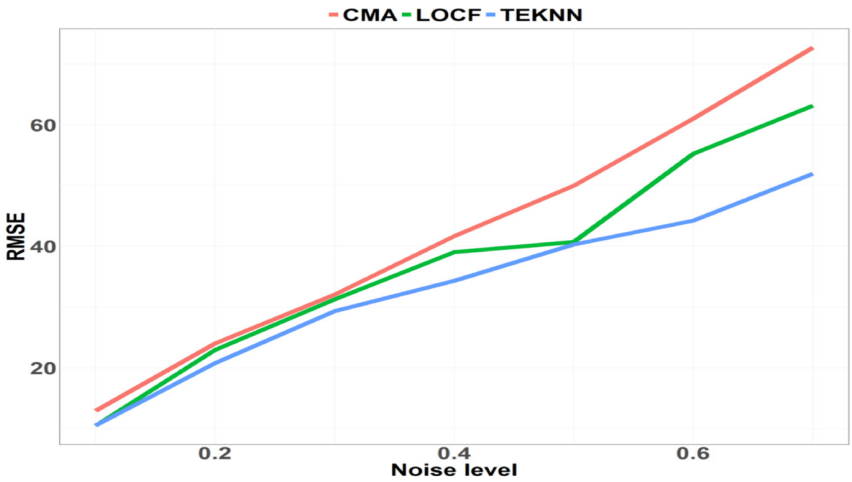


Fig 1. Imputation errors.

https://doi.org/10.1371/journal.pdig.0000115.g001

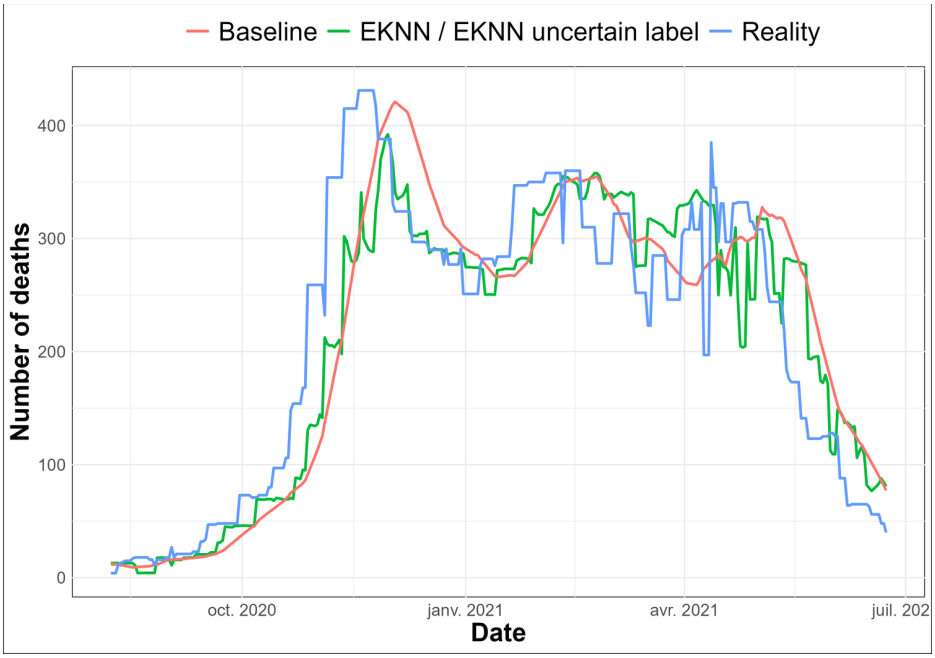


Fig 2. Predictive results on data imputed with time-E*K*NN imputation method, comparison with true labels: $K = 1$: $q = 4$: $v = 0$.

https://doi.org/10.1371/journal.pdig.0000115.g002

are closer to the purple one than the red curve during those periods). During the periods of relative stability when the evolution of the pandemic slows down, there seems to be no significant differences between the considered approaches.

We note a small time shift between the real number of daily deaths and all predictive models, especially at the beginning of the wave. This is due to the fact that the model needs high number of deaths in the past to be able to predict high values in the future.

We observe a phase shift at the start of the wave, due to the fact that, before, there is no neighbor labelled with a high number of deaths, therefore the predicted values are under-estimated until we have data in the training set presenting a high number of deaths. We also note that the phase shift reduces thereafter.

On Fig 3, we observe that, with time-E*K*NN imputations, the "E*K*NN uncertain labels" and the "E*K*NN" make predictions reaching quite similar performances with a slight superiority for the "E*K*NN" (without uncertain model) during the beginning and the end phases of the pandemic wave. During the relatively stable periods, the "E*K*NN" associated with an uncertain model performs better than without imputation uncertain modelling.

The best results with a noise level of zero were obtained with $K = 1$, and $q = 4$. We see on Fig 5 that thanks to the uncertainty model of the TE*K*NN imputation method we have a better predictive performance up to a high noise level. The superiority of the standard E*K*NN after a noise level of about 0.5 is due to the fact that a high missing value rate induces highly uncertain neighborhoods and thus very uncertain predictions, a large mass being attributed to ignorance. The pignistic transformation applied to the mass function output of the E*K*NN distributes the mass on the Ω space in a uniform way on all the singletons; if this mass is too high, the predictions tend to the center of the space. Except for $K = 1$, for all the other configurations $K = \{10, 20\}$ and $q = \{1, 2, \ldots, 7\}$ the use of uncertainty models allows us to have better predictive performances (Figs 4, 5 and 6).

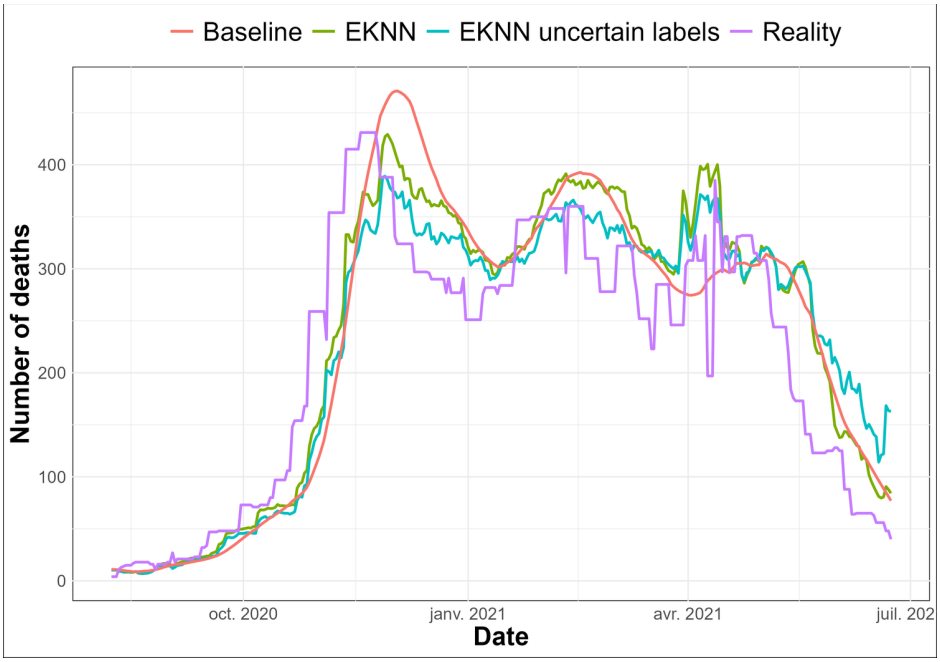


Fig 3. Predictive results on data imputed with time-EKNN imputation method, comparison with true labels: $K = 1$: $q = 3$—$v = 0.3$.

https://doi.org/10.1371/journal.pdig.0000115.g003

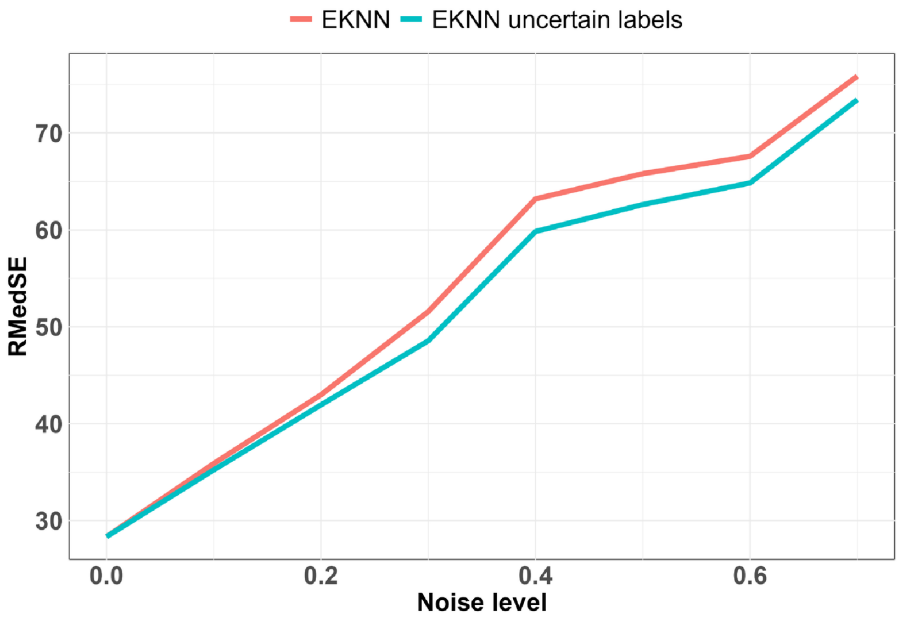


Fig 4. Predictive performances relative to noise levels for data imputed with LOCF imputation method: K = 10: z = 4.

https://doi.org/10.1371/journal.pdig.0000115.g004

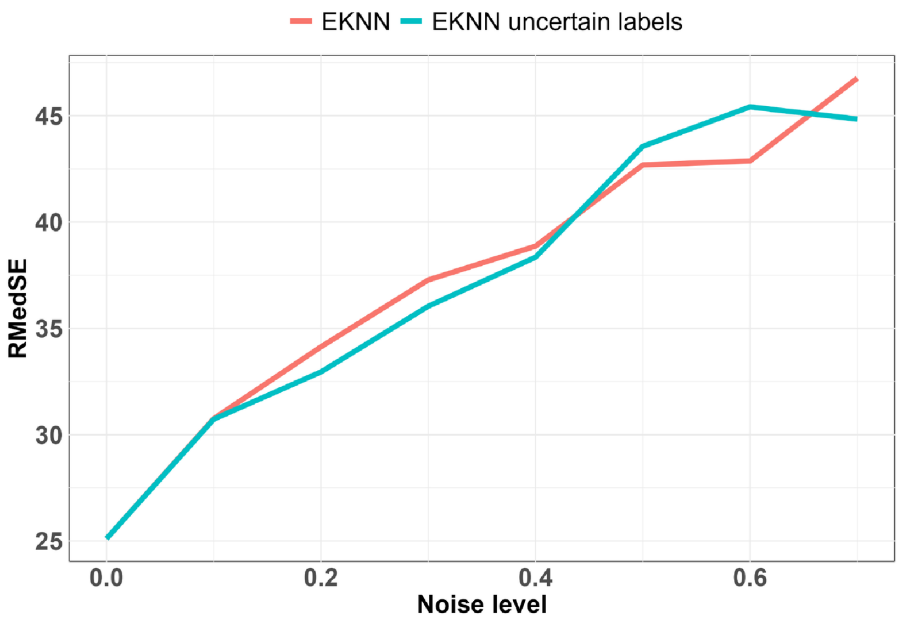


Fig 5. Predictive performances relative to noise levels for data imputed with time-EKNN imputation method: $K = 1$: $q = 4$.

https://doi.org/10.1371/journal.pdig.0000115.g005

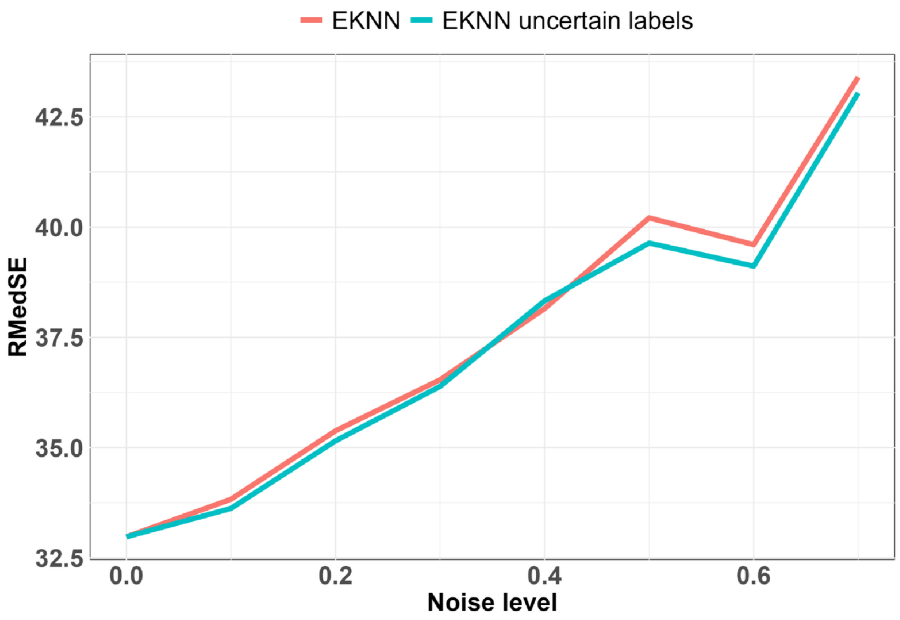


Fig 6. Predictive performances relative to noise levels for data imputed with CMA imputation method: $K = 20$: $q = 4$.

https://doi.org/10.1371/journal.pdig.0000115.g006

Redefining Ω at each iteration depending on the maximum number of deaths observed is a conservative way to proceed, but it allows both models to predict any "reasonable" unobserved value.

3 Conclusion

The aim of this study was to propose uncertainty models associated with missing chronological data imputation methods. The objective was the prediction of the number of daily COVID 19 deaths at a prediction horizon of 7 days with an artificially noised dataset. We proposed three imputation methods (LOCF,CMA,time-E*K*NN) that showed good imputation performances.

For our experiment we extended the E*K*NN methodology proposed in [14] to regression problems. We were able to compare the predictive performances of the "E*K*NN" and the "E*K*NN uncertain labels" with three imputation methods of comparable performances. The experiment showed the benefit of uncertainty modeling for chronological imputed values throughout several hyper-parameters configurations.

The use of incomplete past values $(x_t, y_t)_{t=t-q,\ldots,t}$ as features leads to uncertain feature values. A logical continuation of this work could be to use other predictive models than the E*K*NN, especially the ones that handle uncertain attributes during learning [17, 24, 25].

The problem of predicting COVID 19 daily deaths led us to a numerical regression problem, therefore the time based uncertainty model is not adapted to classification. It would be interesting to extend it to classification in a categorical time series context. We also know from health experts that the number of new COVID 19 cases is not a good indicator for predicting deaths, therefore it would be interesting to weigh the importance of the attributes in the K nearest neighbors computing [26].

Another perspective could be to compare the predictive performance we can get with soft predictive models that handle missing values without any need of imputation.

Additionally, there are some algorithms like E*K*NN that use this framework. The theory of belief functions permits us to have the enhancement of uncertainty modeling as a perspective, for example by using imputation with intervals instead of precise values.

4 Materials and methods

4.1 Background

In this section we expose the basics of belief functions theory, also known as Dempster-Shafer or evidence theory [12, 13] and we detail the Evidential K-Nearest Neighbors algorithm [14].

4.1.1 Belief functions. Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_H\}$ be the so-called frame of discernment, *i.e.* the universe of possible outcomes or hypotheses. The mass function m represents our degree of knowledge about all subsets of Ω, *i.e.* about the powerset 2^Ω of Ω. The elements $A \subseteq \Omega$ such as $m(A) > 0$ are called focal elements and their weights sum to 1:

$$\sum_{A \subseteq \Omega} m(A) = 1 \tag{1}$$

The quantity $m(\Omega)$ represents the degree of ignorance. From the mass function m, different uncertainty measures can be computed such as the *belief* and *plausibility* functions defined in Eqs (2) and (3) which can be interpreted as lower and upper "bounds of probability".

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{2}$$

$$Pl(A) = \sum_{B \cap A \neq \oslash} m(B) \tag{3}$$

Different mass functions can represent different sources of information. At the decision level, it may be necessary to combine them into a single mass function expressing all the knowledge we can infer from these sources.

Mass combination

There are multiple methods of information fusion through mass combination rules [27, 28]. One of the most famous is the Dempster's conjunctive rule of combination $\oplus$ [12] (see Eq (4)):

$$(m_1 \oplus m_2)(A) = \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1(B) m_2(C) \tag{4}$$

where $\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$ is the degree of *conflict* between sources m_1 and m_2.

The main idea of this rule is to consider all sources reliable. After the combination, we get a new mass function that can be used at the decision making level.

Decision making

In cases where the degree of ignorance $m(\Omega)$ is too important, some authors recommend rejecting decision making [14]. Otherwise, the choice of uncertainty measure to make a decision presents a dilemma [29].

For instance, depending on the application goal and the chosen strategy in terms of conservatism, any uncertainty measure lying between the *belief* (2) and the *plausibility* (3) measures can be used. However, those measures are not additive, *i.e.* we do not have $Bel(A \cup B) \neq Bel(A) + Bel(B)\ \forall A, B \in \Omega$ such as $A \cap B = \emptyset$ (same thing for the *Pl* function). For that reason many data science tools are incompatible with those *soft* uncertainty measures since most of them have been developed within the standard probability framework.

For pragmatic reasons many researchers choose to project the information content of mass functions into the standard probability framework [14, 17]. The *Transferable Belief Model* was proposed by Smets [29, 30] where the *pignistic* transformation allows to convert mass functions into standard probability distributions. Despite known drawbacks [31], ignorance degrees are projected into uniform distributions. The pignistic transform defined in Eq (5) remains a natural solution for computing probability distributions from mass functions that mainly relies on uniform ignorance modeling.

In the machine learning context many classifiers make soft predictions expressed in more complex spaces than the standard probability one [14, 32]. When the learning data are uncertain (evidential), in order to get *handy* predictions some authors [17] have proposed to maximize the evidential extension of the likelihood function [14] in order to estimate probability distributions. When the evidential likelihood maximization is not straightforward, the Evidential Expectation Maximization (E^2M) algorithm can be used. However the iterative nature of the E^2M algorithm can lead to high levels of complexity. For the sake of simplicity, the pignistic transform is preferred in this study.

$$BetP(\omega) = \frac{1}{1-m(\emptyset)} \sum_{A \supseteq \omega} \frac{m(A)}{|A|} \quad \forall \omega \in \Omega \tag{5}$$

4.1.2 Evidential *K*Nearest Neighbors—EKNN. The E*K*NN extends *K*-Nearest Neighbors algorithm to the belief functions framework [14] and is based on Dempster's conjunctive rule of combination. Let $(x, y) = (x_i, y_i)_{i=1,\ldots,n}$ be a training set and $\Omega = \{\omega_1, \omega_2, \ldots, \omega_H\}$ the frame of discernment of the class label Y. Let x_s be a new observation to classify. The first step is to compute the distances between x_s and all training examples x_i to get the set of the K "nearest" neighbors of x_s. In the E*K*NN approach, each neighbor is considered as a source of information. For each neighbor x_i labelled with $\{\omega_l\}$, a mass function $m_{s,i}$ is computed:

$$\begin{cases} m_{s,i}(\{\omega_l\}) &= \alpha_0 \times exp(-\gamma_l^2 \times d_{s,i}^{\lambda}) \\ m_{s,i}(\Omega) &= 1 - m_{s,i}(\{\omega_l\}) \end{cases} \tag{6}$$

The quantity $m_{s,i}(\{\omega_l\})$ represents the mass assigned to the label ω_l by neighbor x_i to x_s. The parameters α_0, γ_l can be estimated with classical optimization procedure as gradient descent. The parameter $\gamma_l > 0$ relates to the label ω_l, in [14] the author recommends to set the parameter α_0 (which prevents dogmatic mass functions) to 0.95, $d_{s,i}$ stands for the euclidean distance between x_s and its neighbor x_i and $\lambda \in \{1, 2, 3, \ldots\}$ is a parameter that penalizes the farthest neighbors.

Once all the masses $m_{s,i}$ have been computed, they are combined with Dempster's rule of combination into a final mass $m_s = m_{s,i_1} \oplus \ldots \oplus m_{s,i_K}$ where $m_{s,i_1}, \ldots, m_{s,i_K}$ represent the knowledge associated with the K nearest neighbors of x_s. Finally, decision can be made according to m_s, the approach chosen in [14] is to predict the label corresponding to the maximum of belief.

In the case of uncertain or imperfect labels modeled by mass functions, *i.e.* the learning set is now $(x_i, m_{y_i})_{i=1,\ldots,n}$, for each neighbor x_i we have a mass function m_{y_i} on the label variable Y. In [14], the author proposes to discount the mass functions of all neighbors with the uncertainty level of their labels. In Eq (6), the term corresponding to the uncertainty level of the labels m_{y_i} is added which results in Eq (7).

$$\begin{cases} m_{s,i}(A) &= \alpha_0 \times exp(-\gamma_A^2 \times d_{s,i}^{\lambda}) \times m_{y_i}(A) \quad \forall A \subseteq \Omega \\ m_{s,i}(\Omega) &= 1 - \sum m_{s,i}(A) \end{cases} \tag{7}$$

In this study we deal with a regression problem since the number of COVID-19 daily deaths is numerical. We therefore extend the original E*K*NN model, that was initially designed for classification problems, to discrete regression tasks. To do so we removed all γ_l parameters (defined relatively to categorical class labels) from Eqs (6) and (7), which results in Eqs (8) and (9).

E*K*NN uncertainty model for regression:

$$\text{precise labels} \quad : \quad \begin{cases} m_{s,i}(\{\omega_l\}) &= \alpha_0 \times exp(-d_{s,i}^{\lambda}) \\ m_{s,i}(\Omega) &= 1 - m_{s,i}(\{\omega_l\}) \end{cases} \tag{8}$$

$$\text{uncertain labels} \quad : \quad \begin{cases} m_{s,i}(\{\omega_l\}) &= (\alpha_0 \times exp(-d_{s,i}^{\lambda})) \times m_i(\{\omega_l\}) \\ m_{s,i}(\Omega) &= 1 - m_{s,i}(\{\omega_l\}) \end{cases} \tag{9}$$

The implementation of this extension of the E*K*NN algorithm to regression is available on our github (https://github.com/lgi2p/evidential_imputation).

For decision making (*i.e.* prediction) we used the pignistic transform $BetP_s$ of m_s in order to predict the pignistic expectation.

The predicted label of a new observation x_s is:

$$E_{BetP_s}[Y] = \sum_{\omega \in \Omega} BetP_s(\{\omega\}) \times \omega \tag{10}$$

4.2 Formalism

In this section we present the formalism of both the regression problem and the missing value imputation task.

4.2.1 Predictive problem. Let $D = (x_t, y_t)_{t=0,\ldots,T}$ be a dataset where $x_t \in \Omega_X \subseteq \mathbb{N}$ and $y_t \in \Omega_Y \subseteq \mathbb{N}$ are respectively the feature and label values at time t. We suppose that some label values are missing, *i.e.* some y_t values are not known. The aim of the regression task is to approximate a function f mapping current and past features values to future labels:

$$y_t = f(y_{t-h}, \ldots, y_{t-(h+q)}, x_{t-h}, \ldots, x_{t-(h+q)}) \tag{11}$$

where h is the prediction horizon, q a number of past features and label values to consider. This regression modelling implies that, at any timestamp t, the number of deaths y_t can be predicted from the concatenation of the sets of previous number of death $(y_{t-h}, \ldots, y_{t-(h+q)})$ and of previous number of cases $(x_{t-h}, \ldots, x_{t-(h+q)})$.

Predicting deaths from data restricted to past number of death and cases is not usual in COVID forecasting works since some useful variables as the *basic reproduction number* R_0, hospital entries, exits and intensive care daily numbers, state health measures (confinement, etc) are generally used for predicting future deaths. In our case we chose to restrict to deaths and cases variables as most of the other previously stated variables are usually incomplete in public datasets. Indeed, our work is based on the E*K*NN model which can deal with uncertain labels but not uncertain features (in its initial form).

Moreover, restricting ourselves to only 2 types of data (deaths and cases) makes experiments easier to run. Nevertheless, all this work can be easily extended to high-dimensionality features provided they are not incomplete in dataset. It is worth mentioning that some work has extended the E*K*NN model to uncertain features by computing distances between examples based on Jousselme distance which can be computed between belief function and thus between uncertain features [24].

Since some of D's values are missing, the imputation process has to occur upstream. In Eq (10) past labels y_t are inputs of the function as historic features. Therefore removing incomplete examples introduces irregular timestamps in the data, which is inconsistent with the regularity hypothesis of most time series treatments.

4.2.2 Imputation problem. Let us consider $y_{p_1} \ldots y_{p_U}$ the U known previous label values with $p_U < \ldots < p_1$ before a missing label y_t and $y_{n_1} \ldots y_{n_R}$ with $n_1 < \ldots < n_R$ the next N known values, U and R are hyper-parameters that have to be tuned.

In the example presented in Fig 7 we have $P = 2$ and $N = 3$.

The aim of the imputation process is to compute or *impute* a value $\hat{y}_t$ for all missing y_t.

4.3 Imputation methods

In this section we present three imputation methods to impute $\hat{y}_t$. The first one is the Last Observation Carried Forward (LOCF) method that replaces missing values with the last known value. The second one is the Centered Moving Average imputation (CMA) method that takes into account the dynamical nature of the data, and imputes missing values from the nearest future and past values. The last one is the "time-E*K*NN" (TEKNN) imputation method

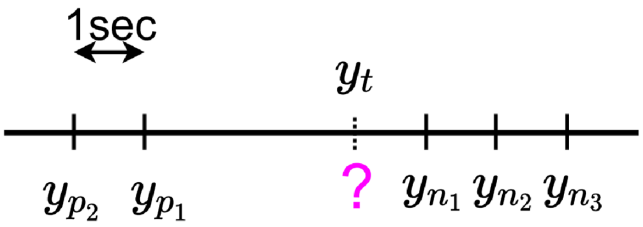


Fig 7. Chronological data imputation.

https://doi.org/10.1371/journal.pdig.0000115.g007

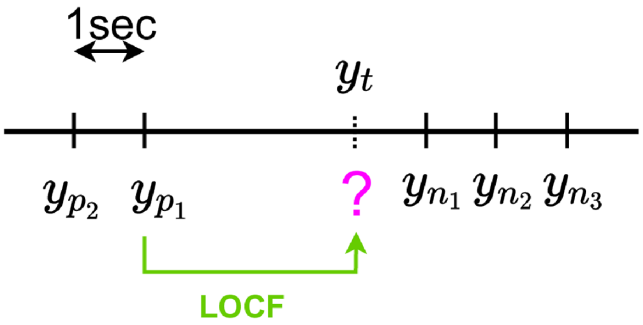


Fig 8. LOCF imputation.

https://doi.org/10.1371/journal.pdig.0000115.g008

that applies the E*K*NN algorithm with a temporal distance to predict missing values, this method also takes into account the dynamical nature of the data as CMA method. The three imputation methods considered in this study are based on the use of those past and future label values.

4.3.1 Last Observation Carried Forward (LOCF). Let y_t be a missing label value and y_{p_1} the last known value. The LOCF imputation is simply $\hat{y}_t = y_{p_1}$ as illustrated in Fig 8.

4.3.2 Centered Moving Average (CMA) imputation. Here we preset a method taking into account the dynamical nature of the data. It is based on the intuition that labels close in time are likely to have close values. More simply, the idea of the CMA imputation method is to impute the missing y_t from the nearest known past and future labels (see Fig 9).

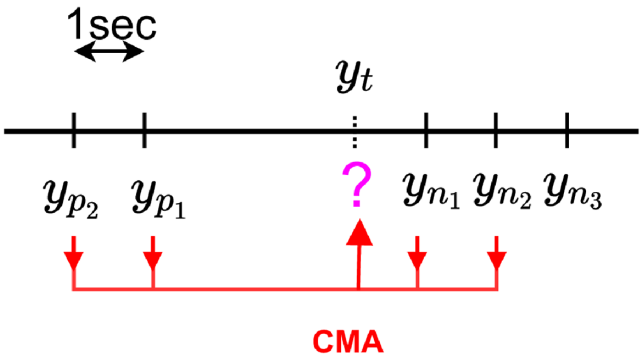


Fig 9. CMA imputation.

https://doi.org/10.1371/journal.pdig.0000115.g009

The past and future label values are averaged according to the duration between them and the missing label y_t. Let $\delta_t^p = |t - p|$ and $\delta_t^n = |t - n|$ be respectively the time shifts between the time t of a missing label y_t and the time of the previous and next known label values, we have:

$$\hat{y}_t = \sum_{u=1}^{U} Sim(t, p_u) \times y_{p_u} + \sum_{r=1}^{R} Sim(t, n_r) \times y_{n_r} \tag{12}$$

with $\forall (t, U, R) \in \{0, ..., T\} \times \mathbb{N}^{*2}$:

- $$Sim(t, p_u) = \frac{sim(t, p_u)}{\sum_{u=1}^{U} sim(t, p_u) + \sum_{r=1}^{R} sim(t, n_r)} \tag{13}$$

- $$sim(t, p_u) = 1 - \frac{\delta_t^{p_u}}{\sum_{u=1}^{U} \delta_t^{p_u} + \sum_{r=1}^{R} \delta_t^{n_r}} \tag{14}$$

Eqs (13) and (14) define a normalized temporal similarity $Sim(t, p_u)$ between a measurement time t and one of the previous measurement times p_u. Those similarities are used to weigh past label values in Eq (12). Note that these equations can be directly transposed to measure the similarity $Sim(t, n_r)$ between t and any next measurement time n_r.

For the example of Fig 7, the CMA imputation with $U = R = 2$ leads to $\hat{y}_t = \frac{6}{30} \times y_{p_2} + \frac{7}{30} \times y_{p_1} + \frac{9}{30} \times y_{n_1} + \frac{8}{30} \times y_{n_2}$

4.3.3 Time-EKNN (TEKNN) imputation. The idea behind this imputation approach is to use the EKNN regression model to predict the missing label values based on the complete

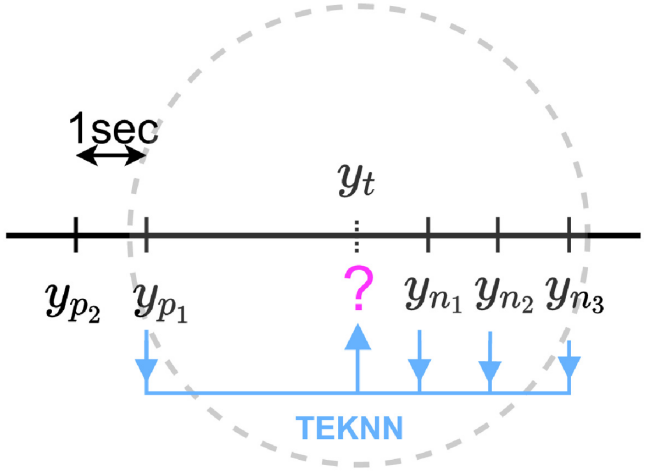


Fig 10. TEKNN imputation.

https://doi.org/10.1371/journal.pdig.0000115.g010

examples, *i.e.* where label values are known, that are the closest in time. This method could be seen as a de-centered extension of the CMA approach where the points used for imputation are the closest regardless of their temporal disposition around the missing value (before and/or after) as in Fig 10. For this model too, the nearest neighbors on time have more weights on the imputed values $\hat{y}_t$ (see Eq (8)).

Regardless of the considered method, by nature the imputation of missing data involves some uncertainty associated with the imputed values. The next subsection proposes an uncertainty model for imputed time series data within the belief function framework.

4.4 Uncertainty modeling for imputation methods

In this subsection we present the uncertainty models associated with the 3 imputation methods described in the previous subsection. After uncertainty modeling, mass functions m_t^y are assigned to each label y_t whether its value is missing or not. For known label values, categorical mass functions $m_t^y(\{y_t\}) = 1$ are assigned. The imputation methods associated with the uncertainty model allow the conversion of a precise but incomplete (in terms of labels) dataset (x, y) into a complete evidential dataset $(x_t, m_t^y)_{t=1,\dots,T}$.

Because of their dynamical nature, the LOCF and CMA imputation methods are associated with a time-based uncertainty model in the rest of this article. The idea behind this is that the closer in time the values used for the imputation of the missing labels are, the less uncertain the resulting imputed labels will be.

The uncertainty model of the time-E*K*NN imputation approach is the E*K*NN's evidential output.

4.5 Time based model

Once missing labels y_t have been imputed with the LOCF or CMA method into precise computed values $\hat{y}_t$, this paragraph describes the evidential uncertainty model associated with the $\hat{y}_t$ values. This modeling aims at discounting or softening these imputed label values according to the duration without available data before and after them. This model is therefore based on the time shifts $\delta_t^p = |t - p|$ and $\delta_t^n = |t - n|$ between the missing values and the closest known ones. The larger those time shifts, the more uncertain the corresponding imputed label $\hat{y}_t$. Let $\beta \in [0, 1]$ be an hyper-parameter controlling the uncertainty level, *i.e.* the decreasing speed of masses $m_t(\{\hat{y}_t\})$ in regards to the time between the missing values and the closest ones. The mass function associated with the CMA and LOCF imputation methods is:

$$\begin{cases} m_t(\{\hat{y}_t\}) &= \exp(-\beta \times \min(\delta_t^{p_1}, \delta_t^{n_1})) \\ m_t(\Omega) &= 1 - m_t(\{\hat{y}_t\}) \end{cases} \tag{15}$$

This model was tested on several experimental set-ups to study its predictive performance.

The overall articulation between imputation and associated uncertainty models is described in Fig 11.

In this section an experiment is presented on a public COVID-19 dataset in which some labels (*i.e.* daily number of deaths) are *noised*, *i.e.* randomly removed and then imputed before learning and testing phases. After describing the dataset, we give the details of our noise procedure and the experimental set-up and finally we analyze the results.

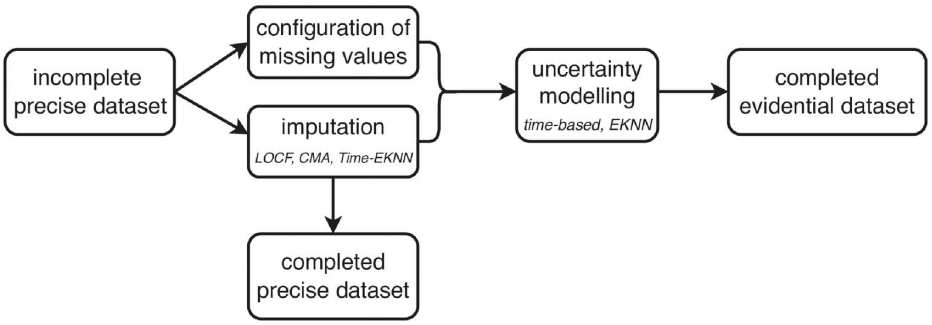


Fig 11. Evidential imputation scheme.

https://doi.org/10.1371/journal.pdig.0000115.g011

4.6 Dataset

On the website ourworldindata (https://ourworldindata.org/) we used the French dataset containing the number of daily confirmed new cases $(x_t)_{t=1,\ldots,T}$ and new deaths $(y_t)_{t=1,\ldots,T}$. Fig 12 shows the evolution of new cases and new deaths.

As the detection policy has evolved between the 2 pandemic waves, the link between new cases and new deaths seems radically different during those 2 periods. As the number of daily new cases was clearly underestimated during the first wave, we restricted the experiment to the second wave. We finally had 367 complete daily observations for this dataset.

As there are no missing values in the dataset, we randomly removed or *noised* some label values (*i.e.* new deaths). In the next subsection we give the details of our noise injection procedure.

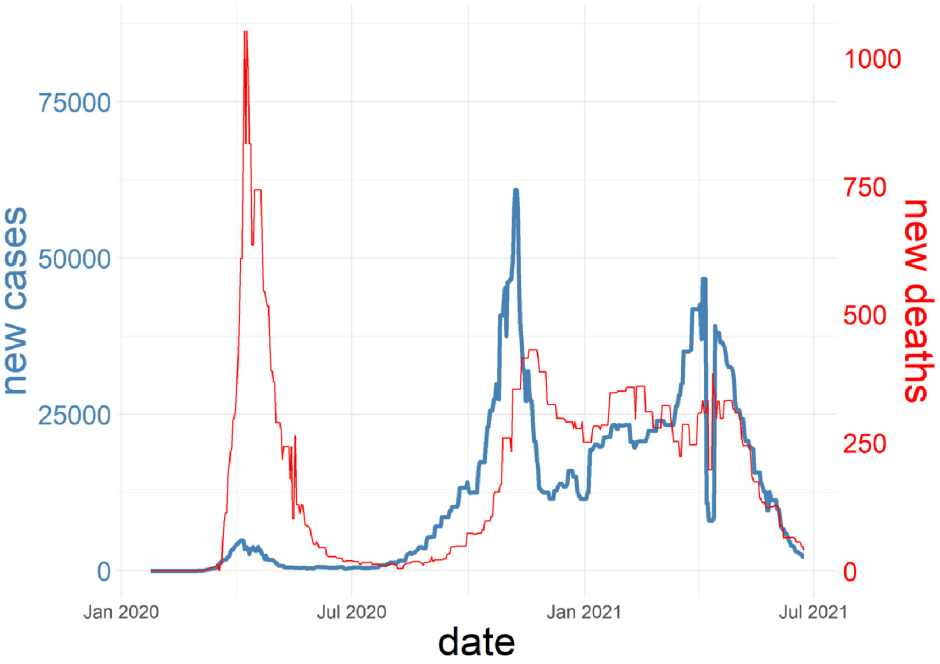


Fig 12. Evolution of new deaths and new cases.

https://doi.org/10.1371/journal.pdig.0000115.g012

4.7 Noise procedure

The proportion $v \in [0, 1]$ of label values y_t to remove is the input of the procedure. In order to simulate plausible measurement errors, we removed labels y_t by time frame. Having randomly picked the frames centers at random, we generated reasonable frame sizes. The procedure is iterative until the proportion of removed labels reaches v.

Algorithm 1: Noise procedure for label values removing.

```
Data: original dataset, v: noise level
Result: noised dataset containing v% of missing labels
removed ← 0
s = ⌊v · T⌋ number of labels to remove;
while removed < s do
  frame center uniform random generation c ∈ [1:T];
  frame size s uniform random generation in ∈ {1, 2, 3};
  computation of label indices to remove {c − s, ....c + s};
  labels removal: {y_{t−s},...,y_{t+s}} → {ŷ_{t−s},...,ŷ_{t+s}};
  removed ← removed + 1 + 2s;
end
```

4.8 Smoothing

Because of the weekly constraint in health policy, the raw COVID-19 data are usually *saw-tooth-shaped* curves. This implies that smoothing method can and should be applied in order to get values corresponding more to the reality than the very noisy raw ones. We chose to use a moving median of 7 days (labels and features). Doing so, we avoided biases without creating unreasonable values. Since smoothed data are usually very regular, imputing missing values on smoothed data is not a real issue, we therefore decided to smooth our noised dataset after imputation. All the implemented predictive models in our study were trained and tested on smooth data because of their higher level of reliability compared to the *sawtooth-shaped* ones. The whole process is represented in Fig 13

4.9 Experimental set-up

In this subsection we present the chronological set-up of our experiment and the considered hyper-parameters spaces. Some hyper-parameters have been set *a priori*: R and U representing the width of the CMA approach (12) were both fixed at 5 days. As hospital reorganization involves strong administrative constraints incompatible with too short or too large horizon, the considered prediction horizon h was 7 days. Finally, the uncertainty hyper-parameter of the time-based uncertainty model (15) β was set at 0.05.

For the other hyper-parameters, several configurations were considered:

- noise level: $v \in \{0, 0.1, \ldots, 0.7\}$

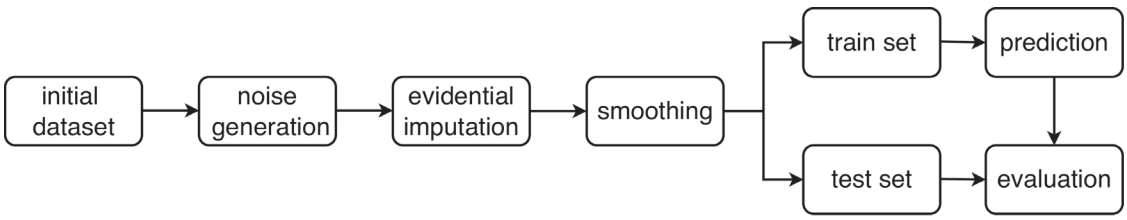


Fig 13. Experimental process.

https://doi.org/10.1371/journal.pdig.0000115.g013

- data historical length: $q \in \{1, 2, \ldots, 7\}$
- number of neighbors for the E*K*NN regressors: $K \in \{1, 10, 20\}$

The data historical length q represents the number of past data (deaths and cases) representing each training example. The first 21 dates are set aside for training, the predictions y_t are then computed iteratively at each date t from all the past couples data $(x_{t'}, y_{t'})_{t'=t-h,\ldots,t-(h+q)}$. At first iteration of the chronological evaluation, the 21 first days are used as training data in order to predict the label value of the $21 + 7 = 28^{th}$ day (with a prediction horizon of $h = 7$ days). After that, the training data are augmented by one date at each iteration, for example at the second iteration we use the 22 first days to predict the label values of the 29^{th} day.

Each complete chronological evaluation is repeated 50 times because of the randomness of the noise procedure and predictions are averaged.

$$(\Omega_Y)_t = \{0, 1, \ldots, \max_{t'<t}(y_{t'}) \times 1.15\} \tag{16}$$

Since Ω_Y must be defined before the prediction step in the regression of the E*K*NN we propose (see Eqs (8), (9) and (10)), we redefine it at each iteration according to Eq 16 by updating the maximum label value in the training data. We chose a safety margin of 15% in regards to the real maximum number of deaths observed. As a baseline we considered the moving average that predicts the number of deaths for the next week as the average of the previous 2 weeks.

Two types of figures are presented. The chronological evaluations (see Figs 2 and 3) allow us to visually evaluate the predictions of the regression for different noise levels according to different imputation methods by comparing the predicted labels with the real ones. E*K*NN regression models are evaluated without and with uncertain models ("E*K*NN" and "E*K*NN uncertain labels"). In the former case, imputed training labels are considered certain whereas in the latter case uncertain models (see Eqs (15) and (8)) allow the E*K*NN regression to take into account data imputation uncertainty.

The noise level sensitivity of the complete evaluations is represented in Figs 6 and 5 where the predictive performance is measured according to the noise level. The evaluation metric we considered is the Root Median Squared Error (RMedSE). We chose it rather than the standard Root Mean Square Error (RMSE) because of the high sensitivity of the mean operator to extreme values which are quite usual in the COVID-19 data.

We evaluate the imputation methods by comparing the initial dataset with the imputed ones in terms of RMSE in order to evaluate the imputation errors that were unlikely to contain extreme values (see Fig 1).

Author Contributions

Conceptualization: Rayane Elimam, Nicolas Sutton-Charani, Jacky Montmain.

Data curation: Rayane Elimam.

Formal analysis: Rayane Elimam, Nicolas Sutton-Charani, Jacky Montmain.

Funding acquisition: Stéphane Perrey.

Investigation: Jacky Montmain.

Methodology: Rayane Elimam, Jacky Montmain.

Project administration: Stéphane Perrey, Jacky Montmain.

Software: Rayane Elimam.

Supervision: Nicolas Sutton-Charani, Stéphane Perrey, Jacky Montmain.

Validation: Nicolas Sutton-Charani, Stéphane Perrey, Jacky Montmain.

Visualization: Nicolas Sutton-Charani.

Writing – original draft: Rayane Elimam.

Writing – review & editing: Nicolas Sutton-Charani, Stéphane Perrey, Jacky Montmain.

References

1. Jerez J., Molina I., García-Laencina P., Alba E., Ribelles N., Martín, M., Franco L. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial Intelligence In Medicine. 2010; 50(2):105–115. https://doi.org/10.1016/j.artmed.2010.05.002 PMID: 20638252
2. Lakshminaryan K., Harp S., Samed T. Imputation of missing data in industrial databases. Applied Intelligence. 1999; 11(3):259–275. https://doi.org/10.1023/A:1008334909089
3. Rubin D. Inference and missing data. Biometrika. 1976; 63(3):581–592. https://doi.org/10.1093/biomet/63.3.581
4. Little R., Rubin D. Statistical analysis with missing data. (Wiley,2002). http://books.google.com/books?id=aYPwAAAAMAAJ.
5. Farhangfar A., Kurgan L., Dy J. Impact of imputation of missing values on classification error for discrete data. Pattern Recognition. 2008; 41(12):3692–3705. https://doi.org/10.1016/j.patcog.2008.05.019
6. Jacquin L., Imoussaten A., Trousset F., Montmain J., Perrin D. Evidential classification of incomplete data via imprecise relabelling: Application to plastic sorting. International Conference On Scalable Uncertainty Management. 2019;122–135. https://doi.org/10.1007/978-3-030-35514-2_10
7. Alizadehsani R., Roshanzamir M., Hussain S., Khosravi A., Koohestani A., Zangooei M., Abdar M., Beykikhoshk A., Shoeibi A., Zare A., Panahiazar M., Nahavandi S., Srinivasan D., Atiya A., Acharya U. Handling of uncertainty in medical data using machine learning and probability theory techniques: A review of 30 years (1991–2020). Annals Of Operations Research. 2021;1–42. https://doi.org/10.1007/s10479-021-04006-2 PMID: 33776178
8. Zadeh L. Fuzzy logic, neural networks, and soft computing. Fuzzy Sets, Fuzzy Logic, And Fuzzy Systems: Selected Papers By Lotfi A Zadeh. 1996;775–782. https://doi.org/10.1142/9789814261302_0040
9. Zadeh L.Fuzzy Sets. Information And Control. 1965; 8:338–353, http://www-bisc.cs.berkeley.edu/Zadeh-1965.pdf. https://doi.org/10.1016/S0019-9958(65)90241-X
10. Dubois D., Prade H. Possibility theory: qualitative and quantitative aspects. Quantified Representation Of Uncertainty And Imprecision,1998; 169–226. https://doi.org/10.1007/978-94-017-1735-9_6
11. Walley, P. Statistical reasoning with imprecise probabilities. (Chapman,1991). ISBN: 0412286602 9780412286605.
12. Dempster A. Upper and lower probabilities induced by a multivalued mapping. Classic Works Of The Dempster-Shafer Theory Of Belief Functions. 2008;57–72. https://doi.org/10.1007/978-3-540-44792-4_3
13. Shafer, G. A mathematical theory of evidence. (Princeton university press,1976)
14. Denœux T. A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Transactions On Systems, Man, And Cybernetics. 1995; 25(5):804–813. https://doi.org/10.1109/21.376493
15. Denœux T. A neural network classifier based on Dempster-Shafer theory. IEEE Transactions On Systems, Man, And Cybernetics-Part A: Systems And Humans. 2001; 30(2):131–150.
16. Elouedi Z., Mellouli K., Smets P. Belief decision trees: theoretical foundations. International Journal Of Approximate Reasoning. 2001; 28(2-3):91–124. https://doi.org/10.1016/S0888-613X(01)00045-7
17. Sutton-Charani, N., Destercke, S., Denœux, T. Learning decision trees from uncertain data with an evidential EM approach. 2013 12th International Conference On Machine Learning And Applications. 2013;1:111-116.
18. Azim, S. & Aggarwal, S. Hybrid model for data imputation: using fuzzy c means and multi layer perceptron. *2014 IEEE International Advance Computing Conference (IACC).* 2014; 1281-1285
19. Li D., Deogun J., Spaulding W., Shuart B. Towards missing data imputation: a study of fuzzy k-means clustering method. International Conference On Rough Sets And Current Trends In Computing. 2004;573–579. https://doi.org/10.1007/978-3-540-25929-9_70

20. Nishanth K., Ravi V. Probabilistic neural network based categorical data imputation. Neurocomputing, 2016; 218:17–25. https://doi.org/10.1016/j.neucom.2016.08.044

21. Hamizadeh J., Moradi M. Enhancing data analysis: uncertainty-resistance method for handling incomplete data. Applied Intelligence. 2020; 50(1):74–86. https://doi.org/10.1007/s10489-019-01514-4

22. Liu, Z., Pan, Q., Mercier, G., Dezert, J. Pattern classification with missing data using belief functions. 17th International Conference On Information Fusion (FUSION 2014;1-8, https://hal-onera.archives-ouvertes.fr/hal-01070496.

23. Jousselme A., Maupin P. Comparison of uncertainty representations for missing data in information retrieval. Proceedings Of The 16th International Conference On Information Fusion. 2013;1902–1909.

24. Trabelsi A., Elouedi Z., Lefevre E. Decision tree classifiers for evidential attribute values and class labels. Fuzzy Sets And Systems. 2019; 366:46–62. https://doi.org/10.1016/j.fss.2018.11.006

25. Tsang S., Kao B., Yip K., Ho W., Lee S. Decision trees for uncertain data. IEEE Transactions On Knowledge And Data Engineering. 2009; 23(1):64–78. https://doi.org/10.1109/TKDE.2009.175

26. Jiao L., Pan Q., Feng X., Yang F. An evidential k-nearest neighbor classification method with weighted attributes. Proceedings Of The 16th International Conference On Information Fusion. 2013;145–150.

27. Florea M., Jousselme A., Bossé, É., Grenier D. Robust combination rules for evidence theory. Information Fusion, 2009; 10(2):183–197. https://doi.org/10.1016/j.inffus.2008.08.007

28. Smets P. Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. International Journal Of Approximate Reasoning, 1993; 9(1):1–35. https://doi.org/10.1016/0888-613X(93)90005-X

29. Smets P., Kennes R. The transferable belief model. Artificial Intelligence, 1994; 66(2):191–234. https://doi.org/10.1016/0004-3702(94)90026-4

30. Smets P. Constructing the Pignistic Probability Function in a Context of Uncertainty. UAI. 1989; 89:29–40.

31. Smets P. Decision making in the TBM: the necessity of the pignistic transformation. International Journal Of Approximate Reasoning. 2005; 38(2):133–147. https://doi.org/10.1016/j.ijar.2004.05.003

32. Yuan B., Yue X., Lv Y., Denœux T. Evidential deep neural networks for uncertain data classification. International Conference On Knowledge Science, Engineering And Management. 2020;427–437.