МЕДИЦИНСКАЯ ИНФОРМАТИКА / MEDICAL INFORMATICS

DOI: https://doi.org/10.60797/BMED.2024.3.2

DATA DISCRETIZATION IN PROGNOSTIC MODELS FOR EPIDEMIOLOGY

Research article

Elistratov S.A.[1, *]
[1] ORCID : 0000-0002-7006-6879;
[1] Sobolev Institute of Mathematics of SB of RAS, Novosibirsk, Russian Federation
[1] Ivannikov Institute for System Programming of RAS, Moscow, Russian Federation

* Corresponding author (sa.elist-ratov[at]yandex.ru)

Abstract

After COVID-19 pandemic, the epidemilogical data prediction had become of a great importance. Since that, numerous different prognostic models, including those involving neural-network based, have been developed, applied and verified. Short-term models are capable to reproduce the oscillacion, but incapable to make a long term prognosis; long-term ones suffer from the noise in the data and require its reduction. In this paper, we propose a method of data prediction using values range discretization as an alternative to the smoothing to get rid of noise-borne problems and applying lag prediction. It is shown that the approach is capable to improve the prognosis quality even for the irregurlar data.

Keywords: epidemiology, neural network, prognosis, discretization.

ДИСКРЕТИЗАЦИЯ ДАННЫХ В ПРОГНОСТИЧЕСКИХ МОДЕЛЯХ ЭПИДЕМИОЛОГИИ

Научная статья

Елистратов С.А.[1, *]
[1] ORCID : 0000-0002-7006-6879;
[1] Институт математики им. С. Л. Соболева СО РАН, Новосибирск, Российская Федерация
[1] Институт системного программирования им. В. П. Иванникова, Москва, Российская Федерация

* Корреспондирующий автор (sa.elist-ratov[at]yandex.ru)

Аннотация

После начала эпидемии COVID-19 важность прогностических моделей для эпидемиологических данных сильно возросла. Благодаря этому разрабатывается, применяется и апробируется множество различных прогностических моделей, включая те, которые основаны на искусственных нейронных сетях. Модели краткосрочного прогноза способны достаточно точно воспроизводить осцилляции, но не способны сделать долгосрочное предсказание; а модели долгосрочного прогноза страдают от статистического шума входных данных и требуют его подавления. В данной работе мы предлагаем прогностический метод, использующий дискретизацию значений в качестве альтернативы сглаживанию с целью шумоподавления и применяющий лаговую модель. Показано, что такой подход позволяет улучшить качество прогноза даже для нерегулярных данных.

Ключевые слова: эпидемиология, нейронные сети, прогноз, дискретизация.

Introduction

Since the COVID-19 pandemic beginning, the demand for forecast in epidemiology greatly increased [1]. The new lethal disease without proven treatment required the estimation of the measures that have had to be taken, as well as a prognosis of the forthcoming ill number. This need pushed the development of the mathematical prognostic models forward. The frequently-reported data as well as the development of the artificial intelligence caused the rise of a large number of neural-network based models, that demonstrate accurate enough results [2], [3], [4], [5].

However, this accuracy turns out to presence for a short-term prognosis (week-length), but is frequently important to know the long-term (month-length) prediction. In long-term prognosis, daily oscillations can be neglected, so just the trend is forecasted. Indeed, oscillations disturb the prognostic model and affects the result. On the other hand, there is an agent-based and cellular automata approach also used in epidemiology [6], [7], [8] that deals with highly discretized data. By this work we propose a method combining discretization and neural-network models, that is capable to yield a long term prediction, and show its efficiency on an example of epidemiological indicators.

Research methods and principles

To make a prognosis, we use a lag model given by a transformation (the graphical scheme of which is depicted on Figure 1):

$$[y_i, y_{i+1}, \ldots, y_{i+(N_{pred}-1)}] = \Phi(y_{i-1}, y_{i-2}, \ldots, y_{i-N_{lag}})$$

As soon as Φ transition function is unknown, it is reasonable to determine it using a neural network as an implicit transition function. To work properly, it should be previously trained on the data available [9]. Fortunately, the data for COVID-19 contains enough records and makes it possible to train the neural network.

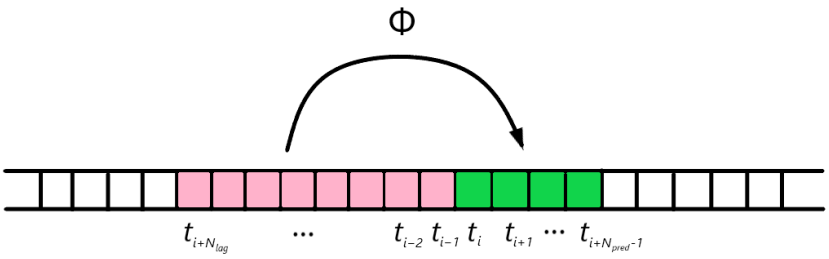

Figure 1 - Scheme of lag prediction
DOI: https://doi.org/10.60797/BMED.2024.3.2.1

The problem arising here is that the data fed into the model for training contains noise that can affect the result. As soon as we are aimed to make a long-term forecast we may neglect it (the reproducing of the noise as accurately as the trend will require too complete model; besides, the exact noise reproduction is hardly necessary, hence it can be believed to be random on the time scales considered). One of the ways to defeat noise disturbance is smoothing [10]. However, instead of smoothing, our approach proposes to reduce oscillations by splitting the values range over several intervals, replacing the data value with a number of the intervals it belongs to. In other words, it can be considered as a replacing data on a temporal interval (x-axis) with the mean value on it, while the interval ends are determined from the uniform vertical slicing (see Figure 2). While selecting the value discretization intervals (y-axis), we shall mind that their length should be so that allow to remove daily oscillations, but small enough to take into account principle data behaviour features.

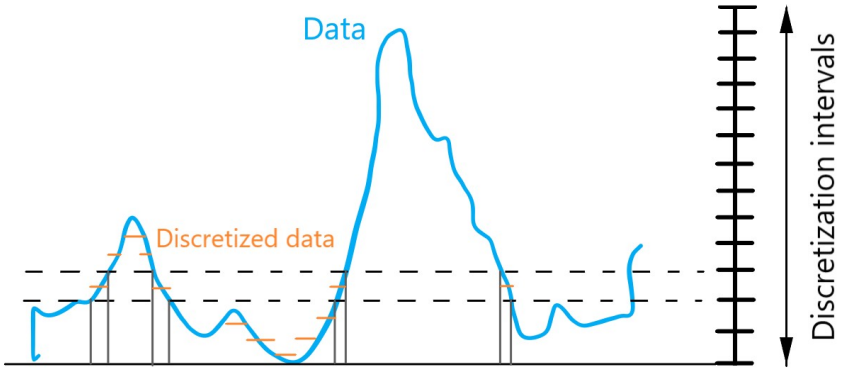

Figure 2 - Scheme of data discretization
DOI: https://doi.org/10.60797/BMED.2024.3.2.2

Ergo, the algorithm proposed consists of the following steps:

1. to obtain a time row of a particular epidemiological indicator (as an example, number of new recoveries over 100.000 people in Moscow);
2. to interpolate the data on a uniform grid (in fact, records may be done over irregular time interval; uniformity is required for the lag mode (see Figure 1));
3. to split the data into train and verification subsets;
4. to discretize the data over values range, as shown on Figure 2;
5. to train the model;
6. to make a prognosis on a validational range and compare with the real data on it.

As an alternative, to compare with, we will use the same scheme without discretization. Instead of point (3) in the list above, we will use EMD-smoothing [10], subtracting several intrinsic modes [11] of the data.

Main results

As to the model, we selected N_{lag} = 20 days, N_{pred} = 10 days and neural network with 3 dense internal layers of 50 units each. The prognosis time was of 50 days. As soon as it is greater than N_{pred}, we made several iterations. The data values were sliced into 50 discrete intervals. For the filtered model, 2 intrinsic modes [11] were subtracted, and log transformation [12] was used.

Figures 3 and 4 demonstrates the prediction results. It is clearly seen that our (discretized) model follows the trend (Figure 4), while the filtered model loose it after 25 days, with the previous prognosis also being not very accurate (Figure 3). Surely, we may complicate and learn further the model for smoothing variant, but the discretized option already yields suitable results.

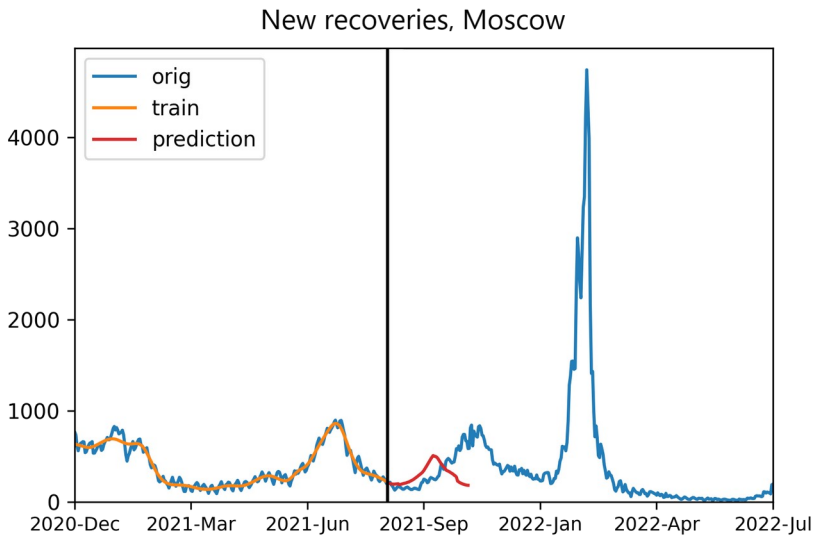

Figure 3 - Prognosis on regular data: smoothing model
DOI: https://doi.org/10.60797/BMED.2024.3.2.3

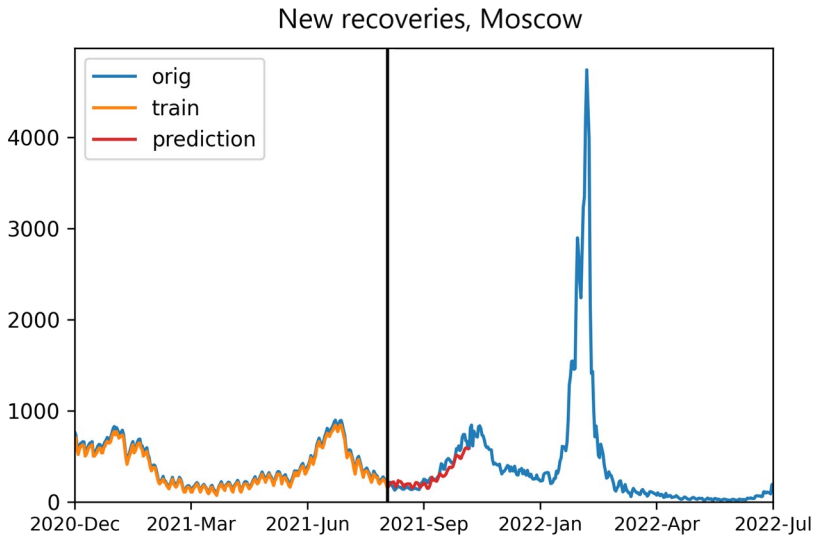

Figure 4 - Prognosis on regular data: discretization model
DOI: https://doi.org/10.60797/BMED.2024.3.2.4

We found that the discretization allows to reach a proper result earlier than the smoothed one do; although, there is another one whilst we consider epidemiological data: the data problem. On Figure 3 one can see a sharp peak near Jan 2022. Epidemiologically, it is connected with a new virus strain appearance, but from the viewpoint of data it means the irregular behaviour. Imagine we have learnt on a previous (regular) data, but the behaviour changed, and we still need the prognosis. Obviously, it would be too presumptuously to expect from the model that it predicts the peak before it starts; however, we can start the prognosis just some time after the peak began (on a slope). The question is, will the model, trained on a regular data and on the very beginning of the peak, reproduce it properly? We may hope it, at least partially, assuming that despite the new strain appearance the principles of the disease spreading remain the same. To test it, we made two predictions, with smoothing and with data range discretization. The results are shown on Figures 5 and 6. The model with discretisation (Figure 5) does not fit the peak exactly, but it does represent at least it duration and the main form. The model with smoothing (Figure 6) fails totally, reproducing neither the duration nor the shape of the peak.

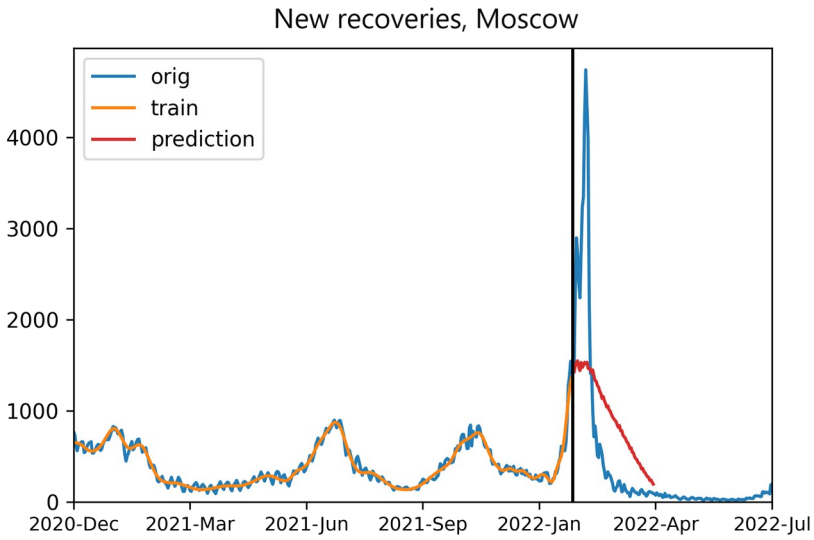

Figure 5 - Prognosis on **irregular** data: smoothing model
DOI: https://doi.org/10.60797/BMED.2024.3.2.5

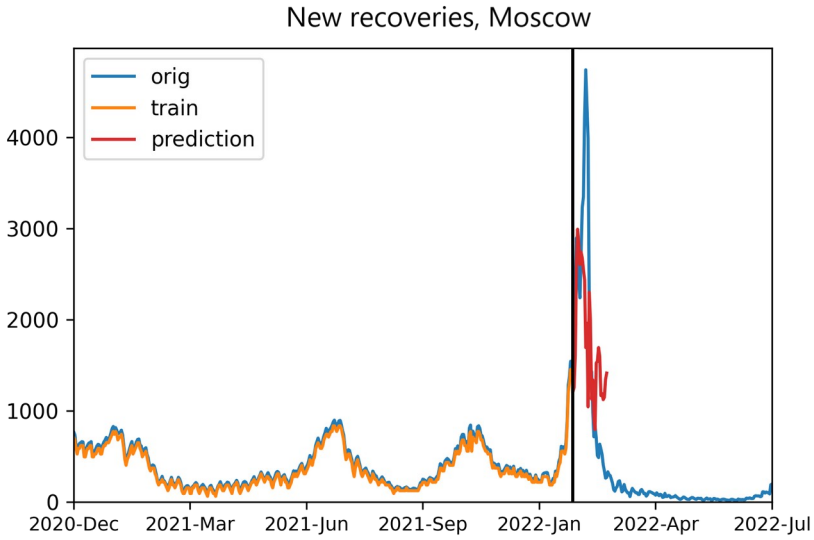

Figure 6 - Prognosis on **irregular** data: discretization model
DOI: https://doi.org/10.60797/BMED.2024.3.2.6

Conclusion

We applied the data range discretization for the lag prediction model with neural network used for an epidemiological indicator forecast. It is shown that the approach increases the accuracy of the prognosis on the same model architecture and can be used for noise reduction instead of the smoothing. The discretization allows accurate prediction also for irregular data, which is important when the prediction required soon after the condition changed (e.g. after a new virus strain appearance). In the contrary, the same model, but with the smoothing used for noise reduction reproduce worse the regular data and makes inadequate forecast for the irregular ones. As a disadvantage, we may consider some data lack during the discretization, which disallows to use the model proposed if further differentiation may be required. Despite this for the pure prediction, for enough number of discretization intervals, it is not critical, as one can see from our forecasting results. We may recommend it as an alternative of smoothing while development forecast models. However, it should be minded that lag predictor works with a neural network, which requires enough data to be trained. These facts may limit the application of the approach proposed (for instance for common diseases like tuberculosis, whose data typically contains monthly records).

Финансирование

Проект 23-71-10068.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Funding

Project 23-71-10068.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы / References

1. Wynants L. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal / L. Wynants, B. Van Calster, G.S. Collins [et al.] // BMJ. — 2020. — Vol. 369. — m1328 p. — DOI: 10.1136/bmj.m1328.

2. Abdulaal A. Prognostic modelling of COVID-19 using artificial intelligence in a UK population / A. Abdulaal, A. Patel, E. Charani [et al.] // Journal of Medical Internet Research. — 2020. — Vol. 22. — № 8. — e20259 p. — DOI: 10.2196/20259.

3. Namadneh N.N. Using artificial neural network with prey predator algorithm for prediction of the COVID-19: The case of Brazil and Mexico / N.N. Namadneh, M. Tahir, W.A. Khan // Mathematics. — 2021. — Vol. 9. — № 2. — 180 p. — DOI: 10.3390/math9020180.

4. Nikparvar B. Spatio-temporal prediction of the covid-19 pandemic in us counties: modeling with a deep LSTM neural network / B. Nikparvar, M.M. Rahman, F. Hatami // Sci Rep. — 2021. — Vol. 11. — № 1. — 21715 p. — DOI: 10.1038/s41598-021-01119-3.

5. Krivorotko O. Modeling of the COVID-19 Epidemic in the Russian Regions Based on Deep Learning / O. Krivorotko, N. Zyatkov // 5th International Conference on Problems of Cybernetics and Informatics. — Baku : IEEE, 2023 — P. 1–5. — DOI: 10.1109/PCI60110.2023.10325993.

6. Fuentes M.A. Cellular automata and epidemiological models with spatial dependence / M.A. Fuentes, M.N. Kuperman // Physica A: Statistical Mechanics and its Applications. — 1999. — Vol. 267. — Issues 3–4. — P. 471–486. — DOI: 10.1016/S0378-4371(99)00027-8.

7. White S.H. Modeling epidemics using cellular automata / S.H. White, A.M. Del Rey, G.R. Sanchez // Appl Math Comput. — 2006. — Vol. 186. — Issue 1. — P. 193–202. — DOI: 10.1016/j.amc.2006.06.126.

8. Krivorotko O. Agent-based modeling of COVID-19 outbreaks for New York state and UK: Parameter identification algorithm / O. Krivorotko, M. Sosnovskaia, I. Vashchenko [et al.] // Infect Dis Model. — 2024. — Vol. 7. — № 1. — P. 30–44. — DOI: 10.1016/j.idm.2021.11.004.

9. Surakhi O. Time-Lag Selection for Time-Series Forecasting Using Neural Network and Heuristic Algorithm / O. Surakhi, M.A. Zaidan, P.L. Fung [et al.] // Electronics. — 2021. — Vol. 10. — № 10. — 2518 p. — DOI: 10.3390/electronics10202518.

10. Елистратов С.А. Пространственное POD-разложение эпидемиологических данных COVID-19 / С.А. Елистратов // Труды Института системного программирования РАН. — 2024. — Т. 36. — № 2. — с. 181–192. — DOI: 10.15514/ISPRAS-2024-36(2)-13.

11. Huang N.E. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis / N.E. Huang, Z. Shen, S.R. Long [et al.] // Proceedings of the Royal Society of London. — 1998. — Vol. 454. — P. 903–995.

12. Feng C. Log-transformation and its implications for data analysis / C. Feng, H. Wang, N. Lu [et al.] // Shanghai Arch Psychiatry. — 2014. — Vol. 26. — № 2. — P. 105–109. — DOI: 10.3969/j.issn.1002-0829.2014.02.009.

Список литературы на английском языке / References in English

1. Wynants L. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal / L. Wynants, B. Van Calster, G.S. Collins [et al.] // BMJ. — 2020. — Vol. 369. — m1328 p. — DOI: 10.1136/bmj.m1328.

2. Abdulaal A. Prognostic modelling of COVID-19 using artificial intelligence in a UK population / A. Abdulaal, A. Patel, E. Charani [et al.] // Journal of Medical Internet Research. — 2020. — Vol. 22. — № 8. — e20259 p. — DOI: 10.2196/20259.

3. Namadneh N.N. Using artificial neural network with prey predator algorithm for prediction of the COVID-19: The case of Brazil and Mexico / N.N. Namadneh, M. Tahir, W.A. Khan // Mathematics. — 2021. — Vol. 9. — № 2. — 180 p. — DOI: 10.3390/math9020180.

4. Nikparvar B. Spatio-temporal prediction of the covid-19 pandemic in us counties: modeling with a deep LSTM neural network / B. Nikparvar, M.M. Rahman, F. Hatami // Sci Rep. — 2021. — Vol. 11. — № 1. — 21715 p. — DOI: 10.1038/s41598-021-01119-3.

5. Krivorotko O. Modeling of the COVID-19 Epidemic in the Russian Regions Based on Deep Learning / O. Krivorotko, N. Zyatkov // 5th International Conference on Problems of Cybernetics and Informatics. — Baku : IEEE, 2023 — P. 1–5. — DOI: 10.1109/PCI60110.2023.10325993.

6. Fuentes M.A. Cellular automata and epidemiological models with spatial dependence / M.A. Fuentes, M.N. Kuperman // Physica A: Statistical Mechanics and its Applications. — 1999. — Vol. 267. — Issues 3–4. — P. 471–486. — DOI: 10.1016/S0378-4371(99)00027-8.

7. White S.H. Modeling epidemics using cellular automata / S.H. White, A.M. Del Rey, G.R. Sanchez // Appl Math Comput. — 2006. — Vol. 186. — Issue 1. — P. 193–202. — DOI: 10.1016/j.amc.2006.06.126.

8. Krivorotko O. Agent-based modeling of COVID-19 outbreaks for New York state and UK: Parameter identification algorithm / O. Krivorotko, M. Sosnovskaia, I. Vashchenko [et al.] // Infect Dis Model. — 2024. — Vol. 7. — № 1. — P. 30–44. — DOI: 10.1016/j.idm.2021.11.004.

9. Surakhi O. Time-Lag Selection for Time-Series Forecasting Using Neural Network and Heuristic Algorithm / O. Surakhi, M.A. Zaidan, P.L. Fung [et al.] // Electronics. — 2021. — Vol. 10. — № 10. — 2518 p. — DOI: 10.3390/electronics10202518.

10. Elistratov S.A. Prostranstvennoe POD-razlozhenie epidemiologicheskih dannyh COVID-19 [COVID-19 Epidemiological Indicators POD Spatial Decomposition] / S.A. Elistratov // Trudy Instituta sistemnogo programmirovanija RAN [Proceedings of the Institute for System Programming of the RAS]. — 2024. — Vol. 36. — № 2. — P. 181–192. — DOI: 10.15514/ISPRAS-2024-36(2)-13. [in Russian]

11. Huang N.E. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis / N.E. Huang, Z. Shen, S.R. Long [et al.] // Proceedings of the Royal Society of London. — 1998. — Vol. 454. — P. 903–995.

12. Feng C. Log-transformation and its implications for data analysis / C. Feng, H. Wang, N. Lu [et al.] // Shanghai Arch Psychiatry. — 2014. — Vol. 26. — № 2. — P. 105–109. — DOI: 10.3969/j.issn.1002-0829.2014.02.009.