International Journal of Biometeorology (2022) 66:2529–2540
https://doi.org/10.1007/s00484-022-02378-z

ORIGINAL PAPER

Prediction of leptospirosis outbreaks by hydroclimatic covariates: a comparative study of statistical models

María José Llop[1,2] · Andrea Gómez[2,3] · Pamela Llop[1,2] · María Soledad López[2,3] · Gabriela V. Müller[2,3]

Received: 25 March 2022 / Revised: 25 August 2022 / Accepted: 30 September 2022 / Published online: 28 October 2022

Abstract

Leptospirosis, the infectious disease caused by a spirochete bacteria, is a major public health problem worldwide. In Argentina, some regions have climatic and geographical characteristics that favor the habitat of bacteria of the *Leptospira* genus, whose survival strongly depends on climatic factors, enhanced by climate change, which increase the problems associated with people's health. In order to have a method to predict leptospirosis cases, in this paper, five time series forecasting methods are compared: two parametric (autoregressive integrated moving average and an alternative one that allows covariates, ARIMA and ARIMAX, respectively), two nonparametric (Nadaraya-Watson Kernel estimator, one and two kernels versions, NW-1 K and NW-2 K), and one semiparametric (semi-functional partial linear regression, SFPLR) method. For this, the number of cases of leptospirosis registered from 2009 to 2020 in three important cities of northeastern Argentina is used, as well as hydroclimatic covariates related to the presence of cases. According to the obtained results, there is no method that improves considerably the rest and can be recommended as a unique tool for leptospirosis prediction. However, in general, the NW-2 K method gets a better performance. This work, in addition to using a long-term high-quality time series, enriches the area of applications of statistical models to epidemiological leptospirosis data by the incorporation of hydroclimatic variables, and it is recommended directing further efforts in this line of research, under the context of current climate change.

Keywords Parametric · Nonparametric · Semiparametric · Leptospirosis outbreak prediction · Hydroclimatic covariates

Introduction

Leptospirosis, the zoonosis caused by the spirochete bacteria *Leptospira interrogans*, is a public health problem all over the world, particularly in tropical and subtropical areas. In Argentina, some regions present climate and geographic characteristics that favor the habitat of the bacteria *L. interrogans*. Infectious diseases, particularly leptospirosis, are climatic-sensitive (Coelho and Massad, 2012; López et al. 2018, 2019; World Health Organization and World Meteorological Organization. Atlas of health and climate, 2012) so that extreme climate events enhanced by climate change increase the problems associated with people's health (Bell et al. 2018; Ebi et al. 2021). For instance, northeastern of Argentina is a region in which extreme precipitation events have increased both, in intensity and frequency, in the last decades (Lovino et al. 2018a and 2018b), favoring the reproduction of the bacteria *L. a interrogans*. This region has important rivers such as Paraná and Uruguay, and the highest precipitation caused significant flooding in the last decade; this trend has continued to rise in recent years (Lovino et al. 2018b). However, since 2019, the region has been under the influence of La Niña period, which has triggered a severe drought, with historical records of low precipitation values and hydrometric levels (Gomes et al. 2021; Naumann et al. 2021). For these reasons, health services must consider, in addition to the treatment of individual cases, the estimated number of cases of such disease also based on the prevailing hydroclimatic conditions. This estimation would improve the response of health systems during potential outbreaks, cutting off, or delaying the disease transmission (Canals, 2010).

✉ María José Llop
llopmariajose@gmail.com

[1] Facultad de Ingeniería Química, Universidad Nacional del Litoral (UNL), Santa Fe, Argentina

[2] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Santa Fe, Argentina

[3] CEVARCAM, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral (UNL), Santa Fe, Argentina

In this direction, this work is attempted to make a statistical study of leptospirosis incidence taking into account the hydroclimatic covariates. Many authors have also remarked that extreme climatic factors, like heavy rainfall and flooding events, increase the incidence of leptospirosis infection (Covertino et al. 2021; Lau et al. 2010; Mwachui et al. 2015). In particular, this study uses a long-term high-quality time series of the leptospirosis incidence and also of the covariates that have been identified in López et al. (2019) as the main hydroclimatic indicators that can influence leptospirosis outbreaks occurrence in northeast Argentina.

Since changes in the covariates may have a time-lagged effect in the leptospirosis incidence, the ecological-environmental process considered in this paper is highly non-linear. As it is well known, in any non-linear system, a first step in the analysis is to identify and quantify the causal relationships between the elements of the system. Following Covertino et al. (2021), that quantification can be made by measuring the information transfer between the elements of the system to infer the presumed causality. This information is measured throughout the transfer entropy (TE), and in particular, in this paper, it would be of interest to compute the TE between covariates and leptospirosis incidence. In addition, in the context of non-linear systems such as leptospirosis and hydroclimatic covariates, to perform a statistical predictive study, it is necessary to determine the optimal time delay that leads a better accuracy. This is made by computing the time delays that minimize the distance between the measures of probabilities of leptospirosis and each covariate or, equivalently, the ones that maximize the mutual information (MI) between them (see Covertino et al. 2021).

On the other hand, to identify the key determinants of leptospirosis variability, a variance-based sensitivity analysis (SA) is performed to see how the variability in the covariates affects the variation in the leptospirosis incidence. SA is widely used in ecological-environmental models, a complete review of the existing methods can be found in Pianosi et al. (2016) (see also Saltelli et al. 2010).

Although in the last years, many mathematical models have been proposed to analyze the spread of infectious diseases, little has been done with leptospirosis. Some references on the deterministic modeling of this disease are Alemneh (2020), Chadsuthi et al. (2021), Gualtieri and Hecht (2019), Warnasekara et al. (2021), and Gómez et al. (2022) and, on the statistically modeling of it, are Cunha et al. (2019), Rahmat et al. (2020), and Souza et al. (2021). Epidemiology modeling can contribute to understanding the transmission characteristics of the disease in particular regions, improving the forecasting performance and, in this way, decreasing the transmission and number of cases.

In the statistical field, the best-known models for time series forecasting are the autoregressive ones. The autoregressive moving average (ARMA) model (Wold, 1938) regresses the response variable in terms of a linear combination of its previous values and various past values of a stochastic term. The autoregressive integrated moving average (ARIMA) model (Box et al. 1994) is a generalization of the ARMA to non-stationary series. The integrated part refers to a differencing initial step, which can be applied to eliminate the non-stationarity of the series. Some applications of this method to epidemiological time series can be found, for instance, in Coutín (2007), Liu et al. (2011), and Promprou et al. (2006). As an extension of ARIMA that allows for covariates, such as hydroclimatic variables, is the so-called ARIMAX method that can be found, for instance, in Kongcharoen and Kruangpradit (2013) and some applications in Chadsuthi et al. (2012) and Desvars et al. (2011).

Concerning to nonparametric methods applied to time series, most of them are based on the classical nonparametric kernel estimator of the regression function, commonly named Nadaraya-Watson (NW) estimator due to its creators (Nadaraya E. 1964 and 1965; Watson, 1964). Nonparametric models are less common in epidemiological studies, and applications of them can be found in the literature mainly for the spatial analysis of leptospirosis. For instance, in Mohammadinia et al. (2019), the authors use support vector machine (SVM) and artificial neural network (ANN) to predict the spatial distribution of leptospirosis. In Cunha et al. (2019), the authors fit nonparametric models to investigate the relationship between the incidence and the explanatory variables. This paper concerns to the kernel estimator introduced by Collomb (1984). In addition, an alternative method that combines both methods developed in Collomb (1984) and Dabo-Niang et al. (2016) is also used. This involves two kernels, one of them controlling the difference between the values of the series and the other one controlling the difference between times.

In 2008, Aneiros-Pérez and Vieu (2008) introduced the semi-functional partial linear regression (SFPLR) model which consists of two terms, one modeling nonparametrically the (temporal) response variable and other adding the additional information presented in the covariates by linearly combining them. This method cuts the long time series trajectory into short curves and uses them as a sample of functional data incorporating to the model one past curve rather than many single past values. This strategy overcomes the problem of choosing or estimating the number of past values to be used in the model. In addition, since the covariates are included in the parametric part of the model, it does not suffer of the curse of dimensionality being, thanks to the nonparametric term, still flexible in terms of model requirements. In the context of infectious diseases with seasonal cycles, this

approach uses the observed value of the covariates only in the month of interest. An important difference with ARIMAX is that SFPLR uses past values of the covariates, which makes it not necessary to predict a future value of them. Although there exists a vast literature concerning with the theoretical study of the SFPRL model, the bibliography concerning to applications is very scarce, especially in the epidemiological area. This model is frequently used to predict electricity demand as in Aneiros et al. (2013), Vilar et al. (2012), and Vilar et al. (2018). The present work aims to enrich the area of applications of these models to epidemiological leptospirosis data by the incorporation of hydroclimatic variables.

The rest of the paper is organized as follows: in the "Data" section, the source of data is presented. The treatment of data and the parameters estimation are explained in the "Numerical implementation and parameter estimation" section. The "Results and discussion" section is devoted to present exploratory, causal, and sensitivity analyses together with the results obtained when applying the prediction methods to leptospirosis. Final conclusions of the work are presented in the "Conclusions and future work" section.

Materials and methods

Data

To perform the analysis, three cities with the highest number of cases of leptospirosis reported in the northeast region of Argentina were selected from the study area (Fig. 1): Santa Fe and Rosario from Santa Fe province and Paraná from Entre Ríos province. Leptospirosis incidence has been recorded in Argentina since 2009, the year in which the National System of Epidemiological Surveillance by Laboratories of Argentina (SIVILA) was implemented. Before this year, there are no reliable records in the country. The analyzed period of time ranges from 2009 to 2020. The confirmed cases of leptospirosis were provided by the Directorate of Health Promotion and Prevention, Ministry of Health of Santa Fe province and by the Epidemiology Division, Ministry of Health of Entre Ríos province. For the mentioned period, the total number of confirmed leptospirosis cases in the three cities was 283: 98 in Santa Fe, 111 in Rosario, and 74 in Paraná. This study does not include suspected, probable, or unconfirmed cases.

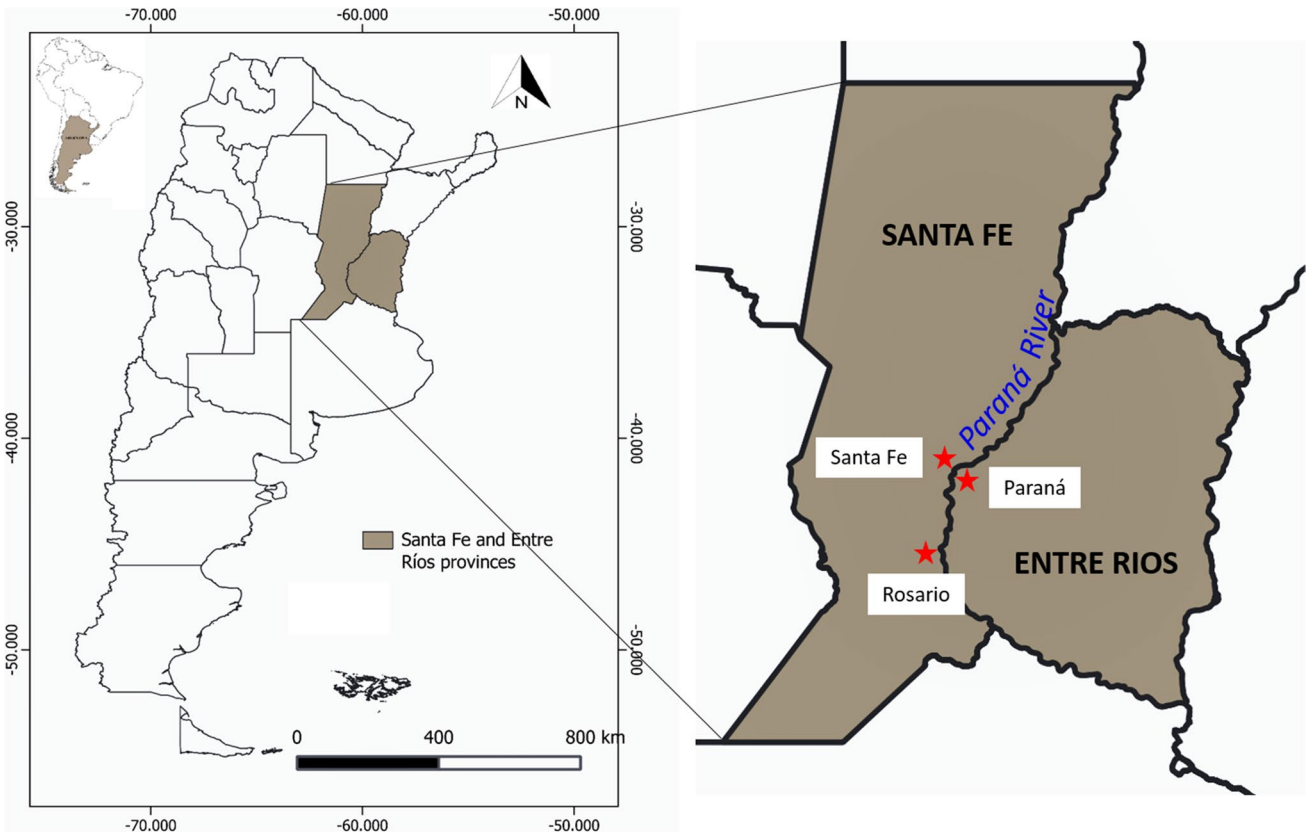

Fig. 1 Left panel: study region in northeast Argentina including Santa Fe and Entre Ríos provinces. Right panel: zoom in of the left panel where the three cities included in the analysis (Santa Fe, Paraná, and Rosario) are marked with red stars

In Fig. 2, the long time series corresponding to the climate covariates and leptospirosis cases for each city are plotted. In the top row of this figure, it can be seen that the number of cases in each city for certain years is small, with many null values and saw-like behavior, which makes any statistical analysis difficult. Unfortunately, as mentioned in the previous paragraph, reliable recorded data for leptospirosis starts in 2009 since, prior to this year, the registration of cases was not mandatory in health centers nor was it systematized or standardized, so there are no reliable records in the region. To overcome this limitation of small sample size, in the "Exploratory analysis" section, the analysis is performed considering the cases of all cities together and increasing, in this way, the number of cases in study.

Selected covariates are those identified in López et al. (2019) as the main hydroclimatic indicators that can influence leptospirosis outbreak occurrence in northeast Argentina. Hydroclimatic datasets include total monthly precipitation, monthly maximum hydrometric river level of the Paraná River, and the Oceanic Niño Index. Precipitation data were provided by the National Weather Service of Argentina (SMN) and the National Institute of Agricultural Technology (INTA). The meteorological stations include Sauce Viejo Aero (which will be called Santa Fe), Paraná, and Rosario Aero. Hydrometric data were provided by the National Water Institute of Argentina (INA). The Oceanic Niño Index (ONI, NOAA/NWS/CPC) is used to determine the years and months under El Niño, La Niña, or neutral conditions. The ONI is the 3-month running mean of the sea surface temperature

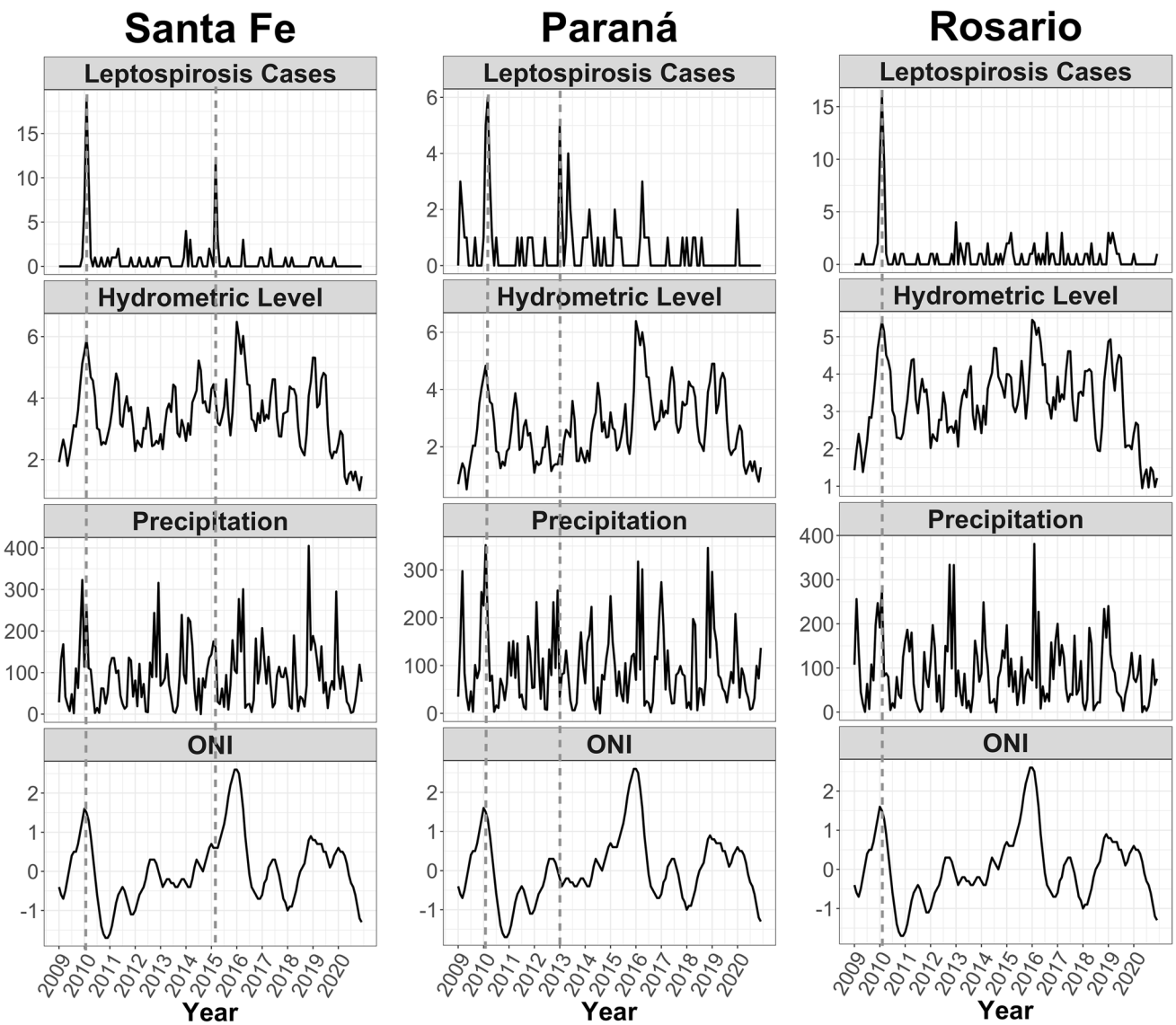

Fig. 2 From top to bottom long time series of leptospirosis cases, monthly maximum hydrometric river level (m), monthly total precipitation (mm), and ONI for Santa Fe, Paraná, and Rosario cities (from left to right). The dashed lines show some correspondence between leptospirosis outbreaks and high values of the covariates

anomaly for the Niño 3.4 region (https://ggweather.com/enso/oni.htm) where the mean is computed with the i-th month, the next month $i+1$, and the past month $i-1$. More precisely, the k-th time lag for the i-th month i is defined as the trimestral mean computed with months $i-k-1$, $i-k$, and $i-k+1$; it is called k-lag. Since the 0-lag involves a future month, it makes no physical sense to analyze it.

Numerical implementation and parameter estimation

The numerical implementation of the prediction methods is performed using the free statistical software R (R Core Team 2013) and its following functions: for ARIMA, arima; for ARIMAX, arimax from package TSA; for Nadaraya-Watson and SFPLR methods, k.smooth from package stats and fregre.plm from package fda.usc, respectively. TE and MI are performed using the OIF Toolbox presented in Li and Covertino, (2021) (see also de webpage https://github.com/HokundaiNexusLab/net-valid). SA is performed using the SAFE Toolbox by Pianosi et al. (2015) (see also de webpage http://bristol.ac.uk/cabot/resources/safe-toolbox/). All the own computational codes are available at https://www.fiq.unl.edu.ar/investigacion/investigacion-reproducible/, and the data (Fig. 2) are available under request.

As usual in statistics, for each city and year (or all cities together), the whole sample is divided into two, the training and the testing samples. The training sample consists of all years except the last one, and it is used to learn the models, this is to estimate the parameters of them. The testing sample consists of the last year (not used in the training sample), and it is used to measure the (out-of-sample) predictive power of the methods. In this direction, the predictive power of each method computed in the testing step, as well as the parameter selection perform in the training step, is measured using the root mean square error (RMSE) given by

$$RMSE = \sqrt{\frac{1}{12}\sum_{i=1}^{12}\left(\hat{Y}_i - Y_i\right)^2},$$

where $\hat{Y}_i$ and Y_i are the predicted and observed monthly values of the last year, respectively. The advantage of this measure of error is that it returns the results in the same units that the original variables, unlikely another commonly used measures like the mean square error.

Both in nonparametric and semiparametric methods, a crucial point is the choice of the measure of closeness between curves. For instance, when the data is smooth, the classical L^2-distance is probably the best choice. However, when the data is rough (as in the case of the data presented in this paper), a more suitable measure of closeness should be used, maybe one that can be used even when the data is not smooth (Ferraty and Vieu, 2006). In this direction, when applying the methods to leptospirosis data, the L^2-distance and the PCA-distance (Ferraty and Vieu, 2006) were compared, obtaining similar results for both metrics; however, given the rough nature of the leptospirosis data (Fig. 3), the PCA-distance is used.

For the nonparametric and semiparametric methods, no assumptions are required to the time series since they are free-model methods (see Collomb 1984; Aneiros-Pérez and Vieu 2008). For the parametric ones, ARIMA and ARIMAX, the methods themselves find the best differencing order to get stationarity and no seasonality, so it is not necessary to check assumptions neither in this case (see Shmueli and Lichtendahl 2016).

Results and discussion

Exploratory analysis

After an exploratory analysis of Fig. 2 where the leptospirosis cases and the covariates are plotted, some correspondence between leptospirosis outbreaks and high values of the covariates for some periods is observed. For instance, in the three cities, at the beginning of 2010, there was an outbreak, and all covariates had high values. In Santa Fe, at the beginning of 2015, the covariates had high values but not extremely high, and the outbreak in that year had fewer cases, which suggests that the covariate values influence the occurrence and magnitude of outbreaks. The same occurs in Paraná in 2013 where the covariates even had lower values. Conversely, in Rosario, at the beginning of 2016, all covariates had high values, but there was not an outbreak, so this could correspond to increased prophylaxis in the city or other variables not considered in this analysis due to lack of information.

The short time series of total cases of leptospirosis, for each city, are plotted in Fig. 3. As can be seen, during outbreak years (2010 for the three cities, 2015 for Santa Fe, and 2013 for Paraná), leptospirosis cases have a certain seasonal behavior in the cities: in all cases, outbreaks occur during the first months of the year. Particularly, in Santa Fe, the highest number of cases is observed between January and April, with outbreaks in February and March. In the rest of the year, the number of cases decreases and remains stable and close to zero. There is a similar behavior in Paraná, where the highest number of cases is observed between January and May, with an outbreak in March. The rest of the year, there are fewer cases. Finally, in Rosario, there is an outbreak in February, and months with a high number of cases extend from January to April, then they decrease, see López et al. (2019) for a deeper analysis of this figure.

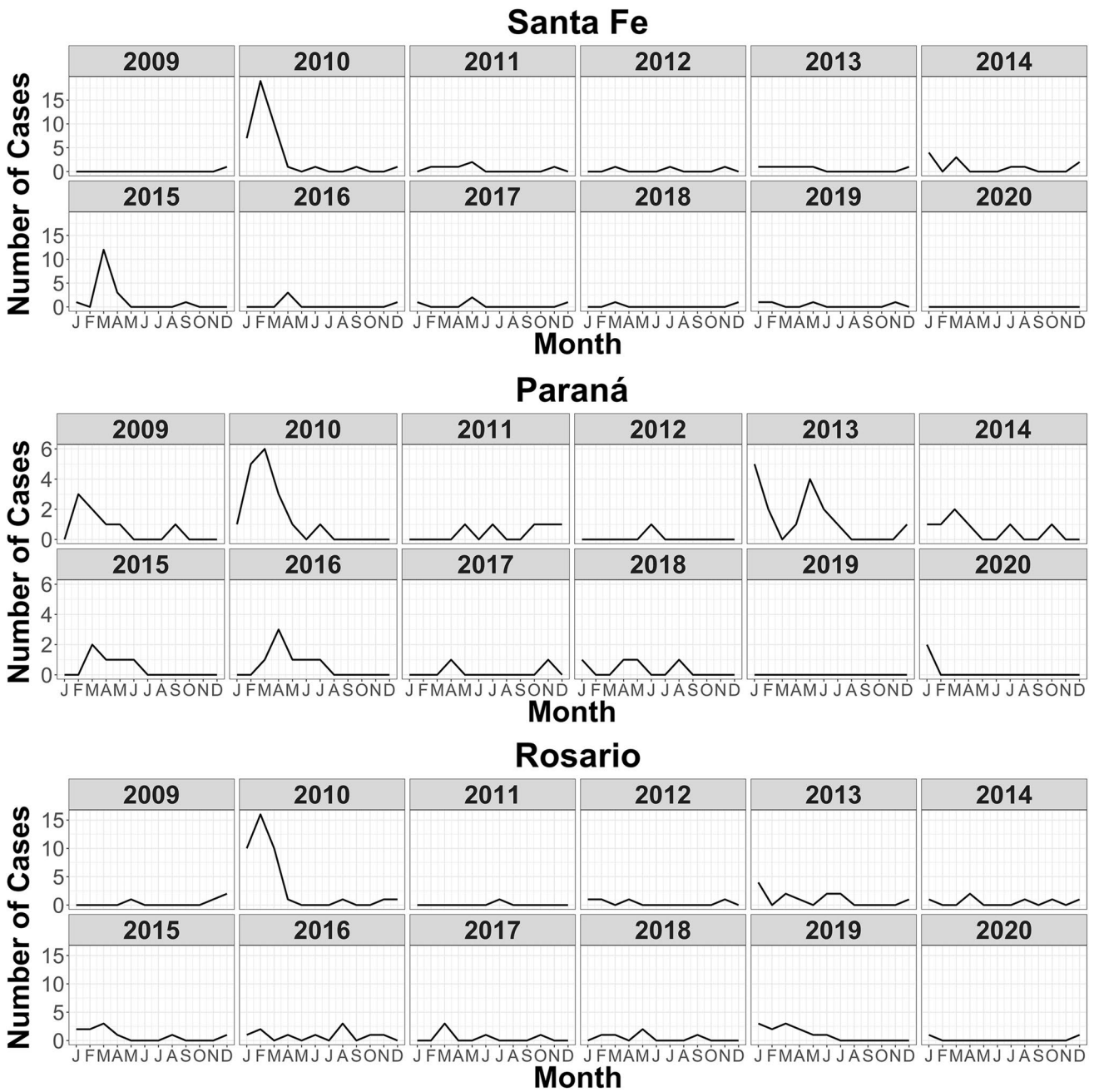

Fig. 3 Short time series of leptospirosis cases in Santa Fe, Paraná, and Rosario cities

Study of causal relationships

To study the causal relationships between covariates and leptospirosis incidence, the TE analysis is performed for all cities, but since the results are similar, for the sake of shortness, just the results for Santa Fe are presented. In this direction, Fig. 4 shows the TE between covariates and leptospirosis incidence for Santa Fe. As can be observed in this figure, there is no significant interaction from leptospirosis to the covariates (black lines) except for small time delays, which, as Covertino et al. (2021) stated, it can be due to numerical mechanism, related to systematic errors that leave overestimation. This indicates that the leptospirosis incidence does not affect neither precipitation, ONI, nor hydrometric level. On the other hand, the reverse effect of the covariates on leptospirosis (gray lines) is detected by the TE, showing a significant causality in that direction (Covertino et al. 2021).

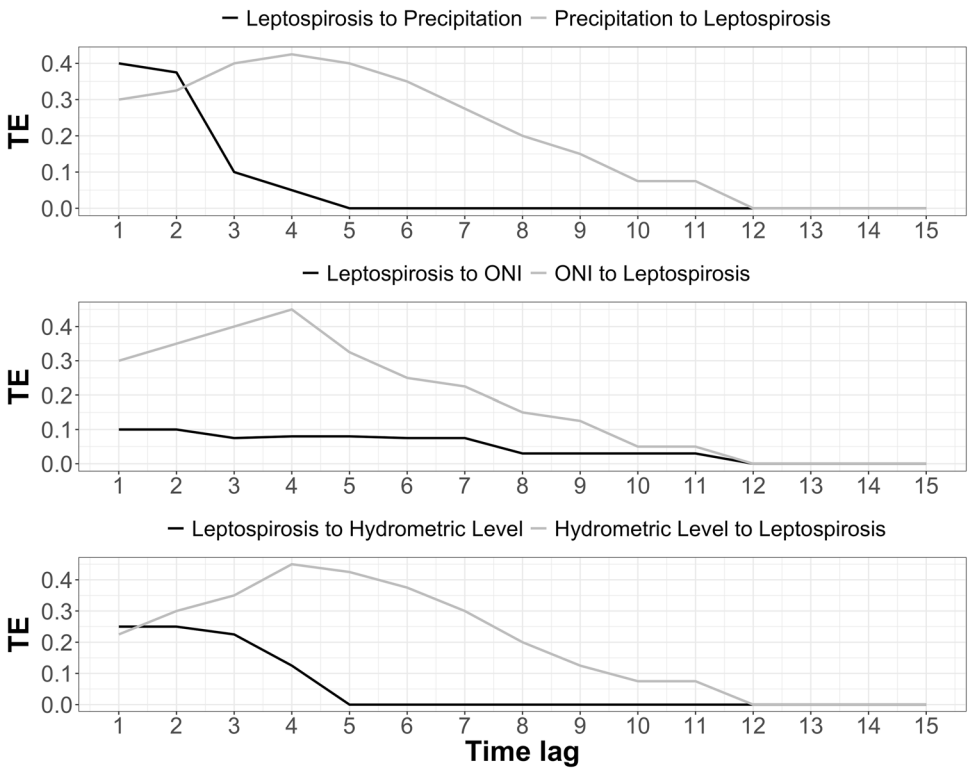

Fig. 4 From top to bottom transfer entropy of covariates and leptospirosis cases for different time delays to Santa Fe city in the study period (2009–2020)

Optimal time delays selection

As was stated in the "Introduction" section, in the context of non-linear systems such as leptospirosis and hydroclimatic covariates, before performing the predictive study, it is necessary to determine the optimal time delay for the covariates. Following Covertino et al. (2021), this is made by choosing the time delay that minimizes the distance between the measures of probabilities of leptospirosis and each covariate or, equivalently, the ones that maximize the MI between them (because it minimizes the uncertainty). In Table 1, the optimal time delays are reported. Those time delays will be used in the five proposed prediction methods applied to predict leptospirosis as a function of the three covariates. For example, to predict leptospirosis cases in 2018 in Santa Fe city, 0, 3, and 4 months delays were used to precipitation, hydrometric level, and ONI index, respectively. As can be observed in the table, in all years in the three evaluated cities, ONI index presents the same or greater delays than the rest of the covariates, as was expected since it is a regional climate indicator unlike precipitations and hydrometric levels which are more local hydroclimatic indicators in each city (Lovino et al. 2018b) and its impact is more immediate in terms of leptospirosis incidence.

Analysis of sensitivity

To perform a variance-based SA, for each specific covariate, the first-order (FI) indices and total order (TI) indices are computed. The FI is defined as the expected reduction in the output (leptospirosis) variability when an input (covariate) is fixed. In consequence, it measures the direct contribution of a covariate to the leptospirosis variance. An FI value near one indicates that the covariate is highly influential in the output variance, whereas an FI near 0 indicates no influence. On the other hand, the TI measures the overall contribution of a covariate considering its direct effect and its interactions with all the other covariates. For one specific covariate, it is defined as 1 minus the expected reduction in the output variance that would be obtained when the rest

Table 1 Optimal time delays for each city, for 2018, 2019, and 2020, in months

	2018			2019			2020		
	Santa Fe	Paraná	Rosario	Santa Fe	Paraná	Rosario	Santa Fe	Paraná	Rosario
Precipitation	0	3	0	0	1	1	0	1	2
Hydrometric level	3	3	0	2	3	0	2	0	0
ONI index	4	4	1	4	3	1	4	3	2

of the covariates are fixed. A TI near zero indicates that the covariate is non-influential (Pianosi et al. 2016). As a result, it was obtained that, for all cities and years, no covariates have individual direct effect to the leptospirosis variance, since the FI obtained in each case was near 0. However, the computed TI (the corresponding table was omitted here since the values are all similar) shows that precipitation and ONI index have a large overall contribution to the leptospirosis variance, since their TI was near 1, following results obtained using other methodologies by López et al. (2019) and Lau et al. (2010).

Prediction

A small number of cases were registered in 2020, the year in which the World Health Organization declared, as a public health emergency of international concern, the pandemic due to the SARS-CoV-2 (COVID-19) virus and the consequent mandatory isolation. Although leptospirosis is a water-borne (not of direct transmission) disease, the apprehension of the population to attend health centers may have resulted in unrecorded leptospirosis cases. In addition, the region has been experiencing a very severe drought since the end of 2019, causing scarce precipitation and hydrometric levels well below normal values (Gomes et al. 2021; Naumann et al. 2021).

As it is the last year of the series and where predictions should be made, and due to this scarcity, the analysis may be not reflecting the real behavior of the long series. In this direction, once the time delays are selected (see the "Optimal time delays selection" section), the five proposed prediction methods were applied to predict leptospirosis as a function of the three covariates, for the last 3 years, 2018, 2019, and 2020. Observed and predicted values are plotted in Fig. 5 for each method, city, and year. Table 2 reports the corresponding RMSE (best results are reported in bold).

It can be observed that, in general, ARIMA and ARIMAX methods have the same behavior in all cities and years, i.e., they predict cases at the beginning of the year, but the predictions are larger than the observed values. This behavior is to be expected since, as can be seen in Fig. 4, the outbreaks occurred at the beginning of the year. This implies that autoregressive methods are sensitive to covariates which present higher values in those months. Nevertheless, there are some particular years in which ARIMAX has good results like Paraná 2018 and 2020 and Santa Fe 2020. With respect to nonparametric methods, although they have an oscillating behavior, they make predictions within the range of the observed values. Consequently, they present better results than the other methods, for example, NW-2 K has the minimum RMSE in Rosario 2018 and 2020, Paraná 2019, and Santa Fe 2018, 2019, and 2020. On the other hand, semiparametric methods have a good performance in some particular years and cities, like Santa Fe 2018 and Paraná 2019, but in general over or underestimate the number of cases.

Note that in Rosario, models without covariates (ARIMA y NW-2 K) have better performance. It can be due to, as is mentioned in the "Data" section, increased prophylaxis or other social variables not considered in this analysis that could influence the probability of infection in Rosario city. Consequently, the covariates used in this study could not be sufficient to explain the behavior of the number of leptospirosis cases. On the other hand, in Santa Fe, semiparametric and nonparametric models have a better performance, whereas in Paraná, ARIMAX has the better performance.

It is observed that, although the results vary in each city and year, in most of them, the nonparametric and semiparametric methods have a better or equal performance than the parametric ones. Parametric methods are more sensitive to outbreak years, which tend to overestimate the number of cases and therefore increase the RMSE, while the semiparametric and nonparametric ones are more adaptive to the shape of the time series through all the years, presenting an oscillating behavior and lowering the RMSE.

In Table 2, it can be also observed that for each city, although the RMSE of all methods does not change considerably from year to year, it could be expected that data (and their distribution) change over time. This shift in the data should be taken into account to perform future predictions using models trained with this data.

All cities together

An analysis merging the data of the three cities together is performed in this section. Due to the scarcity of cases in 2020, as in the individual analysis, a study of the methods in 2018, 2019, and 2020 is carried out. Because the three cities are relatively close (within a distance of no more than 200 km) and present similar precipitation regime, the average of the total monthly precipitation is used as a predictive variable, as well as the ONI, which is a regional variable. The hydrometric level was not considered for this analysis since it has a more local behavior, so the covariates, in this case, are ONI index and precipitation.

Similar to the individual analysis, in order to select the optimal time delays, the MI is minimized (the "Optimal time delays selection" section). In this case, the optimal time lags selected are reported in Table 3. In this case, ONI index presents lower delay than precipitation due to the more regional scale considered.

In Fig. 6, the observed and predicted values are plotted for each method and year. It can be seen that, as mentioned above, 2020 was an atypical year with very few cases concentrated in just two months: three cases in January and

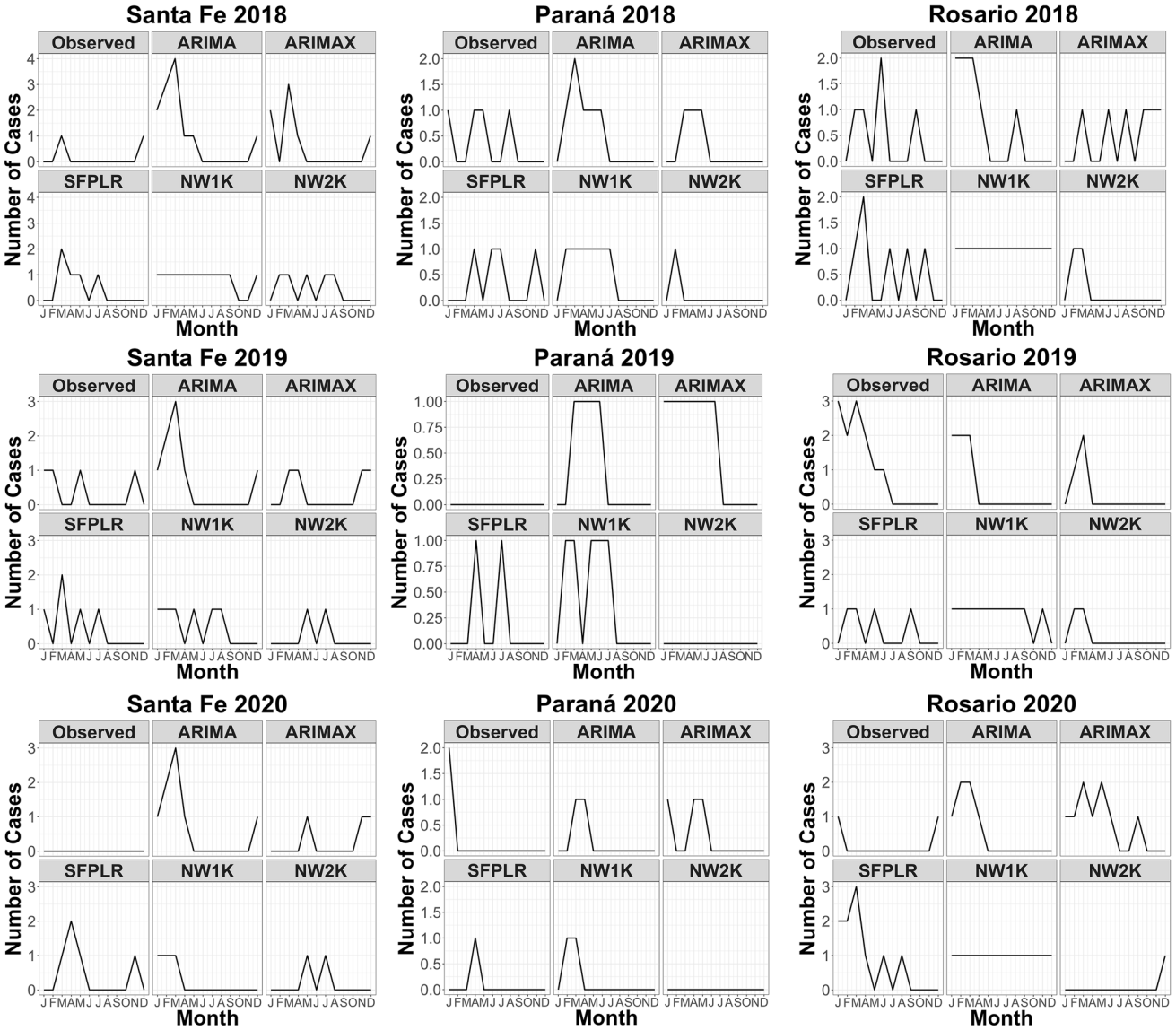

Fig. 5 Observed and predicted values by ARIMA, ARIMAX, NW-1 K, NW-2 K, and SFPLR for Santa Fe, Paraná, and Rosario in 2018, 2019, and 2020

one in December. While in 2018 and 2019, there were more cases. In 2019, there were 19 cases, and most of them were concentrated in the first half of the year. In 2018, there were 11 cases that had a more oscillating behavior throughout the year.

Regarding the predictions, in a similar way to the individual analysis, it can be observed that the parametric methods predict, at the beginning of the year, more cases than those observed. With respect to the semiparametric and nonparametric methods, they predict fewer cases than the

Table 2 RMSE values obtained for 2018, 2019, and 2020 predictions, corresponding to the five methods and the three cities. Best results are reported in bold

	2018			2019			2020		
Method	Santa Fe	Paraná	Rosario	Santa Fe	Paraná	Rosario	Santa Fe	Paraná	Rosario
ARIMA	1.41	0.82	1.04	1.08	0.58	**0.82**	1.15	0.71	0.91
ARIMAX	0.87	**0.50**	0.95	0.71	0.76	1.19	0.50	**0.50**	1.04
NW-1 K	0.82	0.71	0.87	**0.58**	0.65	1.08	0.50	0.71	0.91
NW-2 K	**0.65**	0.65	**0.65**	**0.58**	**0.00**	1.29	**0.41**	0.58	**0.29**
SFPLR	**0.65**	0.71	0.86	0.76	0.41	1.29	0.76	0.65	1.22

Table 3 Optimal time delays for all cities together, for 2018, 2019, and 2020

	2018	2019	2020
Precipitation	2	2	2
ONI index	1	1	1

Table 4 RMSE values obtained for 2018, 2019, and 2020 predictions, corresponding to the five methods and the data of the three cities merged. Best results are reported in bold

	2018	2019	2020
ARIMA	1.41	1.32	1.50
ARIMAX	1.29	1.47	1.47
NW-1 K	1.19	1.63	1.22
NW-2 K	**0.87**	**1.26**	**1.08**
SFPLR	1.55	1.66	1.66

parametric ones, which correspond more to the observed values, although the predictions are oscillating and do not show a specific pattern. Table 4 shows the values of the RMSE obtained with each method (best results are reported in bold). Although the parametric methods seem to adapt to the behavior of the time series (predicting more cases at the beginning of the year), the predicted values are overestimating the values observed, and consequently, the RMSE is greater for those methods. While NW-2 K has the lower value of RMSE in all years.

Conclusions and future work

In this work, the performance of five time series prediction methods was compared. In a first instance, two parametric methods were implemented: the classical ARIMA model and a new alternative that allows for covariates (ARIMAX). Then, two nonparametric methods were performed: the classical Nadaraya-Watson kernel estimator and an extension that involves two kernels, one controlling the differences between the value of the series and the other controlling the distances between times. Finally, the performance of a semiparametric method, the SFPLR was implemented.

These methods were applied to predict leptospirosis cases that occurred during 2018, 2019, and 2020 years in three cities of northeast Argentina, Santa Fe, Paraná, and Rosario. The statistical analysis was carried out in each city separately, and since the number of registered cases in each city for certain years is small and presents many null values, it was also carried out merging the data of all cities together.

Previous to the statistical analysis, a causal relationship study between covariates and leptospirosis and also an optimal time delay selection for any covariate were performed. The former analysis was carried out analyzing the TE and the last one by maximizing the MI.

According to the obtained prediction results, there is no unique method that improves considerably the rest and can be recommended as a tool for leptospirosis prediction. However, in general, nonparametric methods (without covariates) got a better performance than the others, particularly the NW-2 K. This result is observed both, in the analysis of all the cities together and in the individual analysis. On the other side, parametric methods are more sensitive to outbreaks and tended to overestimate the number of cases, while the nonparametric ones presented an oscillating behavior in the range of observed values, which decreased the RMSE. The methods that use covariates (ARIMAX and SPFLR) are not yet able of reliably capturing the observed relationship between leptospirosis outbreaks and hydroclimatic variables as it was expected, and this can be attributed to the nature of the data.

To improve the performance of the methods in the future, the number of years of data should be increased by systematically recording confirmed cases of the disease in time by the public health authorities. Incorporate other covariates to the methods in addition to the climatic ones that are related to leptospirosis also could improve its performance,

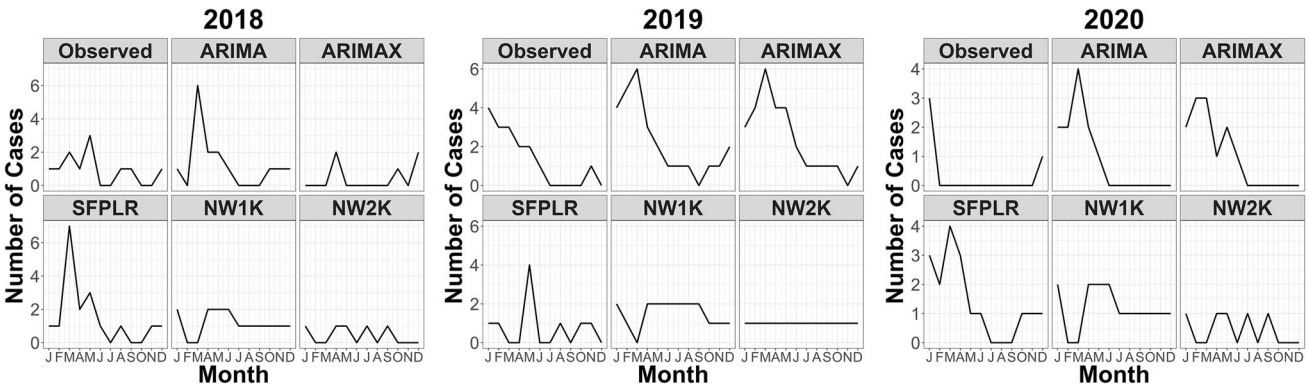

Fig. 6 Observed and predicted values by ARIMA, ARIMAX, NW-1 K, NW-2 K, and SFPLR for all cities together in 2018, 2019 and 2020

for example, social factors, such as health system prophylaxis that might influence the transmission rate of infection (López et al. 2019; Resende Londe et al. 2016).

Despite the aforementioned aspects, this work, in addition to using a long-term high-quality time series, enriches the area of applications of statistical models to epidemiological leptospirosis data by the incorporation of hydroclimatic variables. Also, it can be said that predictive models of this climate-sensitive disease could become useful early warning tools in health systems, in the context of current climate change, directing more efforts in this line of research.

As mentioned in the "Prediction" section, shifts in the data over time could generate model performance degradation, so in a future work, it could be of interest to analyze how the incidence distribution changes as a function of the predictors variables. There exists different kind of shifts that can occur depending on which variable changes its distribution over time and which method it will be applied. For instance, for methods without covariates, just the distribution of the incidence should be monitored over time. However, for methods including covariates, it is important to analyze how shifts in the covariates affect the leptospiroris distribution.

Acknowledgements The authors would like to thank the Directorate of Health Promotion and Prevention, Ministry of Health of Santa Fe province, the Epidemiology Division, Ministry of Health of Entre Ríos province, the National Weather Service of Argentina (SMN), the National Institute of Agricultural Technology (INTA), and the National Water Institute of Argentina (INA).

Declarations

Competing interests The authors declare no competing interests.

References

Alemneh H (2020) A co-infection model of dengue and leptospirosis diseases. Adv Differ Equ 1:1–23

Aneiros G, Vilar J, Cao R, Muñoz San Roque A (2013) Functional prediction for the residual demand in electricity spot markets. IEEE Trans Power Syst 28(4):4201–4208

Aneiros-Pérez G, Vieu P (2008) Nonparametric time series prediction: a semi-functional partial linear modelling. J Multivar Anal 99(5):834–857

Bell J, Langford Brown C, Conlon K, Herring S, Kunkel K, Lawrimore J, Luber G, Schreck C, Smith A, Uejio C (2018) Changes in extreme events and the potential impacts on human health. J Air Waste Manag Assoc 68(4):265–287

Box G, Jenkins G, Reinsel G (1994) Time series analysis: forecasting and control, 3rd edn. Prentiuce Hall Canada

Canals M (2010) Short-term predictability of influenza ah1n1 cases based on deterministic models. Rev Chilena Infectol 27(2):119–125

Chadsuthi S, Chalvet-Monfray K, Wiratsudakul A, Modchang C (2021) The effects of flooding and weather conditions on leptospirosis transmission in thailand. Sci Rep 11(1):1486

Chadsuthi S, Modchang C, Lenbury Y, Iamsirithaworn S, Triampo W (2012) Modeling seasonal leptospirosis transmission and its association with rainfall and temperature in thailand using time-series and arimax analyses. Asian Pac J Trop Med 5:539–546

Coelho M, Massad E (2012) The impact of climate on leptospirosis in São Paulo, Brazil. Int J Biometeorol 56:233–241

Collomb G (1984) "Propriétés de convergence presque compléte du prédicteur á noyau". Z. Wahrschein-lichkeitstheorie verw Gebiete 66:441–460

Coutín M (2007) Utilización de modelos arima para la vigilancia de enfermedades transmisibles en Cuba. Revista Cubana de Salud Pública 33(1). https://www.redalyc.org/articulo.oa?id=21433212

Covertino M, Reddy A, Liu Y, Muñoz-Zanzi C (2021) Eco-epidemiological scaling of leptospirosis: vulnerability mapping and early warning forecasts. Sci Total Environ 799, [149102]. https://doi.org/10.1016/j.scitotenv.2021.149102

Cunha M, Costa F, Ribeiro G, Carvalho M, Reis R, Júnior N, Pischel L, Gouveia E, Santos A, Queiroz A, Wunder E, Reis M, Diggle P, Ko A (2019) Rainfall and other meteorological factors as drivers of urban transmission of leptospirosis. PLOS Negl Trop Dis 16(4):e0007507

Dabo-Niang S, Ternynck C, Yao A (2016) Nonparametric prediction of spatial multivariate data. J Nonparametr Stat 28(2):428–458

Desvars A, Jégo S, Chiroleu F, Bourhy P, Cardinale E, Michault A (2011) Seasonality of human leptospirosis in reunion Island (Indian Ocean) and its association with meteorological data. PLoS ONE 6(5):1–10

Ebi KL, Vanos J, Baldwin J, Bell J, Hondula D, Errett N, Hayes K, Reid C, Saha S, Spector J, Berry P (2021) Extreme weather and climate change: population health and health system implications. Annu Rev Public Health 42(1):293–315

Ferraty F, Vieu P (2006) Nonparametric functional data analysis. Theory and Practice. Springer, New York

Gomes MS, Cavalcanti IFA, Müller GV (2021) Drought impacts on south america and atmospheric and oceanic influences. Weather Clim Extremes 34(4):100404

Gómez A, López MS, Müller G, López L, Sione W, Giovanini L (2022) Modeling of leptospirosis outbreaks in relation to hydroclimatic variables in the northeast of Argentina. Heliyon 8(6):e09758

Gualtieri A, Hecht J (2019) An epidemic model for the propagation of leptospirosis outbreaks. J Health Sci 7:135–141

Kongcharoen C, and Kruangpradit T (2013) Autoregressive integrated moving average with explanatory variable (arimax) model for Thailand export. Conference: the 33rd International Symposium on Forecasting

Lau CL, Smythe LD, Craig SB, Weinstein P (2010) Climate change, flooding, urbanisation and leptospirosis: fuelling the fire? Trans R Soc Trop Med Hyg 104:631–638

Li J, Covertino M (2021) Inferring ecosystem networks as information flows. Sci Rep 11(7094):1–22

Liu Q, Liu X, Jiang B, Yang W (2011) Forecasting incidence of hemorrhagic fever with renal syndrome in china using arima model. BMC Infect Dis 11(1):218

López M, Müller G, Sione W (2018) Analysis of the spatial distribution of scientific publications regarding vector-borne diseases related to climate variability in South America. Spat Spatio-Temporal Epidemiol 26:35–93

López M, Müller G, Lovino M, Gómez A, Sione E, Aragonés Pomares L (2019) Spatio-temporal analysis of leptospirosis incidence and its relationship with hydroclimatic indicators in northeastern Argentina. Sci Total Environ 694(4):133651

Lovino M, Müller O, Berbery E, Müller G (2018a) How have daily climate extremes changed in the recent past over northeastern Argentina? Global Planet Change 168:78–97

Lovino M, Müller O, Müller G, Sgroi L, Baethgen W (2018b) Interannual-to-multidecadal hydroclimate variability and its sectoral

impacts in northeastern Argentina. Hydrol Earth Syst Sci Discuss 22:3155–3174
Mohammadinia A, Saeidian B, Pradhan B, Ghaemi Z (2019) Prediction mapping of human leptospirosis using ann, gwr, svm and glm approaches. BMC Infect Dis 19(2):917
Mwachui MA, Crump L, Hartskeerl R, Zinsstag J, Hattendorf J (2015) Environmental and behavioural determinants of leptospirosis transmission: a systematic review. PLOS Negl Trop Dis 9(9):e0003843
Nadaraya E (1964) On estimating regression. Theory of Probability & Its Applications 9:141–142
Nadaraya E (1965) On nonparametric estimates for density functions and regression curves. Theory of Probability & Its Applications 10:297–302
Naumann G, Podesta G, Marengo J, Luterbacher J, Bavera D, Arias Muñoz C, Barbosa P, Cammalleri C, Chamorro L, Cuartas A, de Jager A, Escobar C, Hidalgo C, Leal de Moraes O, N M, Maetens W, Magni D, Masante D, Mazzeschi (2021) The 2019–2021 extreme drought episode in La Plata basin. EUR 30833 EN, Publications Office of the European Union, Luxembourg. https://doi.org/10.2760/773
Pianosi F, Beven K, Freer J, Hall J, Rougier J, Stephenson DB, Wagener T (2016) Sensitivity analysis of environmental models: a systematic review with practical workflow. Environ Model Softw 79:214–232
Pianosi F, Sarrazin F, Wagener T (2015) A Matlab toolbox for global sensitivity analysis. Environ Model Softw 70:80–85
Promprou S, Jaroensutasinee M, Jaroensutasinee K (2006) Forecasting dengue haemorrhagic fever cases in Southern Thailand using ARIMA models. WHO Regional Office for South-East Asia. Dengue Bull 30:99–106
Rahmat F, Zulkafli Z, Juraiza Ishak A, Mohd Noor S, Yahaya H, Masrani A (2020) Exploratory data analysis and artificial neural network for prediction of leptospirosis occurrence in seremban, malaysia based on meteorological data. Front Earth Sci 8:377
ResendeLonde L, Silva da Conceição R, Bernardes T, de Assis Carvalho, Dias M (2016) Flood-related leptospirosis outbreaks in Brazil: perspectives for a joint monitoring by health services and disaster monitoring centers. Nat Hazards 84:1419–1435
Saltelli A, Annoni P, Azzini I, Campolongo F, Ratto M, Tarantola S (2010) Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. Comput Phys Commun 181:259–270
Souza K, Góes J, Melo M, Leite P, Andrade L, Goes M, Nunes Ribeiro C, Araújo D, Menezes A, Santos A (2021) Spatiotemporal clustering, social inequities and the risk of leptospirosis in an endemic area of brazil: a retrospective spatial modelling. Trans R Soc Trop Med Hyg 115(8):854–862
R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/
Shmueli G, Lichtendahl KC (2016) Practical time series forecasting with R: A hands-on guide, 2nd edn. Axelrod Schnall Publishers, Florida
Vilar J, Aneiros G, Raña P (2018) Prediction intervals for electricity demand and price using functional data. Int J Electr Power Energy Syst 96:457–472
Vilar J, Cao R, Aneiros G (2012) Forecasting next-day electricity demand and price using nonparametric functional methods. Int J Electr Power Energy Syst 39(1):48–55
Warnasekara J, Agampodi S, Rupika Abeynayake R (2021) Time series models for prediction of leptospirosis in different climate zones in Sri Lanka. PLoS ONE 16(5):1–18
Watson G (1964) Smooth regression analysis. Sankhya Series A 26:359–372
Wold H (1938) A study in the analysis of stationary time series, vol 102. Almqvist and Wiksells Boktryckert Uppsala, London
World Health Organization and World Meteorological Organization (2012)Atlas of health and climate. World Health Organ