Clinical Microbiology and Infection 26 (2020) 123.e1–123.e7

Contents lists available at ScienceDirect

Clinical Microbiology and Infection

journal homepage: www.clinicalmicrobiologyandinfection.com

Original article

A prospective prediction tool for understanding Crimean–Congo haemorrhagic fever dynamics in Turkey

Ç. Ak [1], Ö. Ergönül [2], M. Gönen [3, 4, *]

[1] Graduate School of Sciences and Engineering, Koç University, İstanbul, Turkey
[2] Department of Infectious Diseases and Clinical Microbiology, School of Medicine, Koç University, İstanbul, Turkey
[3] Department of Industrial Engineering, College of Engineering, Koç University, İstanbul, Turkey
[4] School of Medicine, Koç University, İstanbul, Turkey

ARTICLE INFO

Article history:
Received 23 October 2018
Received in revised form
18 April 2019
Accepted 9 May 2019
Available online 24 May 2019

Editor: Dr. M Leeflang

Keywords:
Crimean–Congo haemorrhagic fever
Gaussian processes
Machine learning
Spatiotemporal epidemiology
Vector-borne disease

ABSTRACT

Objectives: We aimed to develop a prospective prediction tool on Crimean–Congo haemorrhagic fever (CCHF) to identify geographic regions at risk. The tool could support public health decision-makers in implementation of an effective control strategy in a timely manner.
Methods: We used monthly surveillance data between 2004 and 2015 to predict case counts between 2016 and 2017 prospectively. The Turkish nationwide surveillance data set collected by the Ministry of Health contained 10 411 confirmed CCHF cases. We collected potential explanatory covariates about climate, land use, and animal and human populations at risk to capture spatiotemporal transmission dynamics. We developed a structured Gaussian process algorithm and prospectively tested this tool predicting the future year's cases given past years' cases.
Results: We predicted the annual cases in 2016 and 2017 as 438 and 341, whereas the observed cases were 432 and 343, respectively. Pearson's correlation coefficient and normalized root mean squared error values for 2016 and 2017 predictions were (0.83; 0.58) and (0.87; 0.52), respectively. The most important covariates were found to be the number of settlements with fewer than 25 000 inhabitants, latitude, longitude and potential evapotranspiration (evaporation and transpiration).
Conclusions: Main driving factors of CCHF dynamics were human population at risk in rural areas, geographical dependency and climate effect on ticks. Our model was able to prospectively predict the numbers of CCHF cases. Our proof-of-concept study also provided insight for understanding possible mechanisms of infectious diseases and found important directions for practice and policy to combat against emerging infectious diseases. **Ç. Ak, Clin Microbiol Infect 2020;26:123.e1–123.e7**
© 2019 European Society of Clinical Microbiology and Infectious Diseases. Published by Elsevier Ltd. All rights reserved.

Introduction

Crimean–Congo haemorrhagic fever (CCHF) is a tick-borne viral infection usually transmitted by tick bites, or through contact with tissues, blood or other bodily fluids from infected people and animals [1]. Turkey has the highest case counts among other countries where it remains endemic. *Hyalomma marginatum* ticks are the primary vectors, and they feed on animals at each developmental stage. Both wild and domesticated animals are important in the disease transmission cycle, serving as reservoirs for the continuation of tick re-infection.

People working or living close to livestock or to habitats of the vector ticks are particularly at risk. Human-to-human transmission is possible, typically among health-care workers or care-givers. When the possibility for enzootic transmission exposure increases, the risk of CCHF virus infection for humans increases as well [2]. Environmental changes can influence both the survival and reproduction of *H. marginatum* ticks, then may trigger community outbreaks. For example, neglect of agricultural lands and agricultural reforms causing landscape alterations may be an important factor for the emergence of CCHF. The investigation of those environmental factors that may influence the cycle of CCHF is relevant for outbreak preparedness and response.

* Corresponding author. Mehmet Gönen, College of Engineering, Koç University, Rumelifeneri Yolu, 34450 Sarıyer, İstanbul, Turkey.
E-mail address: mehmetgonen@ku.edu.tr (M. Gönen).

https://doi.org/10.1016/j.cmi.2019.05.006
1198-743X/© 2019 European Society of Clinical Microbiology and Infectious Diseases. Published by Elsevier Ltd. All rights reserved.

Some of the seasonal and climatic covariates were previously reported as important predictors of CCHF virus infections [3–5]. Areas with higher temperatures, precipitation and humidity were linked with high CCHF occurrence in Bulgaria and Iran [4,5]. Suitable habitat for *H. marginatum* ticks was reported as fragmented agricultural lands, forested lands and grass cover in Turkey and Bulgaria, and non-irrigated agricultural land (e.g. pasture) was found to be correlated with CCHF case counts in Turkey [5–7].

The use of spatiotemporal modelling tools might help us better understand the characteristics of established outbreaks to develop different types of interventions to prevent and treat diseases. Predicting the emergence is not realistic because there are so many variables; nevertheless predicting the spatial and temporal trajectory is feasible and probably more effective [8]. Such studies were carried out for Ebola, Zika, H1N1 influenza, and severe acute respiratory syndrome viruses and the results of these studies helped decision-makers to plan bed capacity [9], anticipate travel-related spread [10] and plan vaccine trials [11].

World-wide CCHF retrospective risk maps were reported using the published cases [12], however, a prospective risk analysis based on a comprehensive set of data including climatic, environmental and husbandry parameters is still lacking. Turkey has the highest number of laboratory-confirmed CCHF cases. Monthly data covering 14 years and comprising >10 000 cases could be valuable for understanding the spatiotemporal dynamics of disease spread. We have already presented the improved performance of a structured Gaussian process (GP), against frequently used machine-learning algorithms used in ecological and epidemiological applications [13]. Here we describe the spatiotemporal dynamics of CCHF and extract the important covariates for CCHF virus infection using a structured GP method on the surveillance data set for Turkey. We tested the generalization capability of our approach by predicting where and how many CCHF cases will be observed in each month in 2016 and 2017 prospectively.

Methods

The surveillance data consist of monthly case counts (i.e. observations) for each province. Our regression model takes the past case counts and covariate information as inputs and outputs a numeric value as the future case count.

Surveillance data

The date (i.e. month and year) and location (i.e. province) of the laboratory-confirmed CCHF cases in Turkey between January 2004 and December 2015 were obtained from the Ministry of Health to train our predictive model. We were provided with the surveillance data between January 2016 and December 2017 after we made our predictions for those years. In our study, the province centres were used as the case locations.

Agricultural, demographic, geographic, meteorological and temporal covariates

We collected over 50 potentially related spatial and temporal covariates for use as input in our model. These covariates are listed in the Supplementary material (Table S1). Detailed interpretations of the covariates are presented at http://midas.ku.edu.tr/ProspectiveCCHF.

Latitudes, longitudes and altitudes of province centres were taken from the website of the General Directorate of Highways (http://www.kgm.gov.tr). The remaining spatial covariates were obtained from the Census of Agriculture Agricultural Holdings (Households) of Turkey, which can be found on the website of the Turkish Statistical Institute (http://www.turkstat.gov.tr). Year and month information was extracted from the surveillance data given. CCHF cases had been observed frequently during hot months (e.g. May, June and July), moderately during warm months (e.g. April, August and September) and rarely during cold months (e.g. October, November, December, January, February and March). We encoded each time period by three temporal covariates: the year, month and seasonal group (i.e. hot, warm or cold) to which it belonged.

Climate covariates were taken from the Climatic Research Unit database [14], and other temporal covariates were obtained from the website of the Turkish Statistical Institute. The number of households was divided by the total population of each province and land-related covariates were divided by the total area of each province to make these covariates comparable across different provinces.

Gaussian processes

Gaussian process regression is a machine-learning algorithm that finds a relation between an output y (e.g. CCHF cases) and a set of inputs x (e.g. longitude, latitude, date, etc.). The main assumption of this model is that there is an unobserved or latent function f that depends on x, but for which we only have access the version with some noise, y. This unobserved variable is a GP with the mean vector μ and covariance matrix Σ, which depends on the inputs [15]. In this study, we formulated a GP model with a Kronecker decomposition approach for spatiotemporal modelling, named structured GP, to learn covariance functions for both knowledge extraction and prediction. Our main hypothesis about the spatiotemporal processes is that response values depend on both time and location. We need a kernel function (i.e. covariance function) that makes nearby observations in time and/or space produce similar values. Each spatial and temporal covariate is fed into a kernel function for structured GP, (see Supplementary material, Appendix S1, for a detailed description).

We get a better understanding about the underlying dynamics of the process to be modelled when data can be explained with fewer covariates, which may be hidden or latent factors that in combination play greater roles in the observed dynamics. To find these fewer but important covariates, we optimized each covariate's relative importance.

For the 2016 prediction, we used the years 2004–2015 as training sets (81 provinces $\times$ 144 months). We then used the trained model to predict case counts of 81 provinces for 2016 (81 provinces $\times$ 12 months). For the 2017 prediction, we used the years 2004–2016 as training sets (81 provinces $\times$ 156 months). We then used the trained model to predict case counts of 81 provinces for 2017 (81 provinces $\times$ 12 months).

A study in Turkey found that areas with CCHF cases had lower mean temperatures in the late autumn and the winter [16]. We used the fact that vector-borne disease dynamics are affected by the previous year's weather conditions, animal population, etc. because vector abundance is also affected by these. Hence, covariates of this year will be used to make predictions for the case counts of next year. We trained our model using all spatial covariates, the temporal covariates between 2003 and 2014 and the case counts for years 2004–2015; then given all the spatial covariates and temporal covariates of the year 2015 and the learned parameters from our trained model we predicted the cases for 2016. The same approach was applied for the 2017 predictions. We focused on prospective predictions of the years 2016 and 2017. Prediction for any given year can be done given the covariates of the previous year.

The Pearson's correlation coefficient and normalized root mean squared error were used to measure the prediction performance.

Computational modelling was performed using the statistical software package R [17].

Source codes

The input covariates, nationwide CCHF surveillance data set and our computational results reported in this study can be publicly explored and downloaded at http://midas.ku.edu.tr/ProspectiveCCHF/.

Results

Spatial and temporal distribution of cases

In Turkey, 10 411 confirmed CCHF cases were reported between years 2004 and 2017, mainly from April to October, and yearly epidemic curves peaked around June and July (Fig. 1a). Most of these confirmed CCHF cases were reported in north and northeast regions of Anatolia (Fig. 1b). Detailed interpretations of the case counts are presented at http://midas.ku.edu.tr/ProspectiveCCHF/.

Prospective prediction for 2016 and 2017

We predicted the nationwide annual case count for 2016 as 438, whereas the observed case count was 432 (Fig. 2). Similarly, we predicted the nationwide annual case count for 2017 as 341, whereas the observed case count was 343 (Fig. 3). Pearson's correlation coefficient and normalized root mean squared error values for the 2016 prediction scenario are 0.83 and 0.58, respectively. For the 2017 prediction, Pearson's correlation coefficient is 0.87 and normalized root mean squared error is 0.52. Each month's prediction for all provinces on a map can be seen at http://midas.ku.edu.tr/ProspectiveCCHF/.

Covariate importance

Latitude and number of settlements with <25 000 inhabitants covariates of provinces (i.e. spatial covariates) and monthly potential evapotranspiration (evaporation and transpiration) measurements (i.e. temporal covariate) were found to be the most explanatory covariates for the 2016 prediction (see Supplementary material, Figure S1a,b). In the 2017 prediction, number of settlements with <25 000 inhabitants and longitude covariates of provinces (i.e. spatial covariates) and monthly potential evapotranspiration measurements (i.e. temporal covariate) were the most important covariates (see Supplementary material, Fig. S1c,d).

Discussion

Turkey has the highest number of laboratory-confirmed CCHF cases, and we included all 10 441 CCHF cases into our computational analyses. We used a unified model including a rich collection

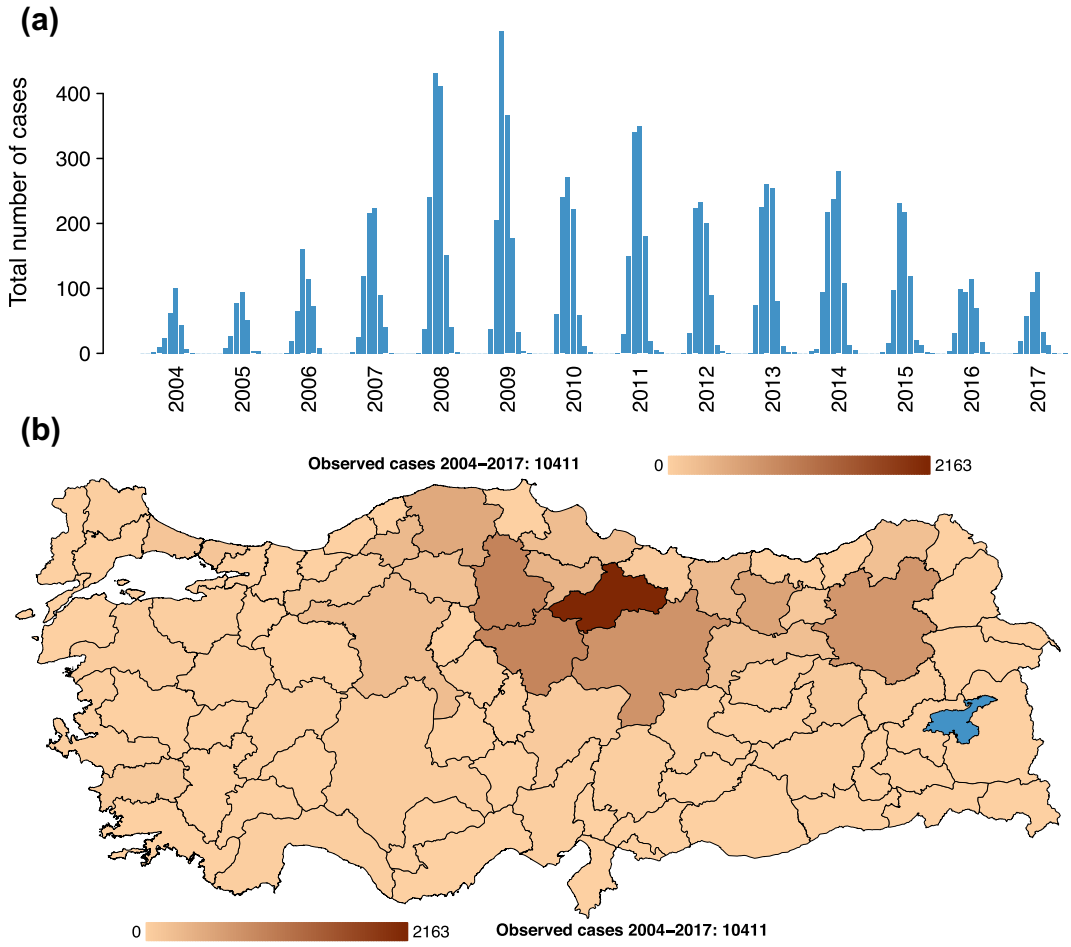


Fig. 1. Summary of Turkish nationwide Crimean–Congo haemorrhagic fever (CCHF) surveillance data set. (a) Monthly confirmed CCHF case counts between January 2004 and December 2017. (b) Total confirmed CCHF case counts for each province between years 2004 and 2017. Numbers in the key of (b) correspond to the minimum and maximum numbers of observed cases in provinces between 2004 and 2017. Yearly case count maps can be seen at http://midas.ku.edu.tr/ProspectiveCCHF/.

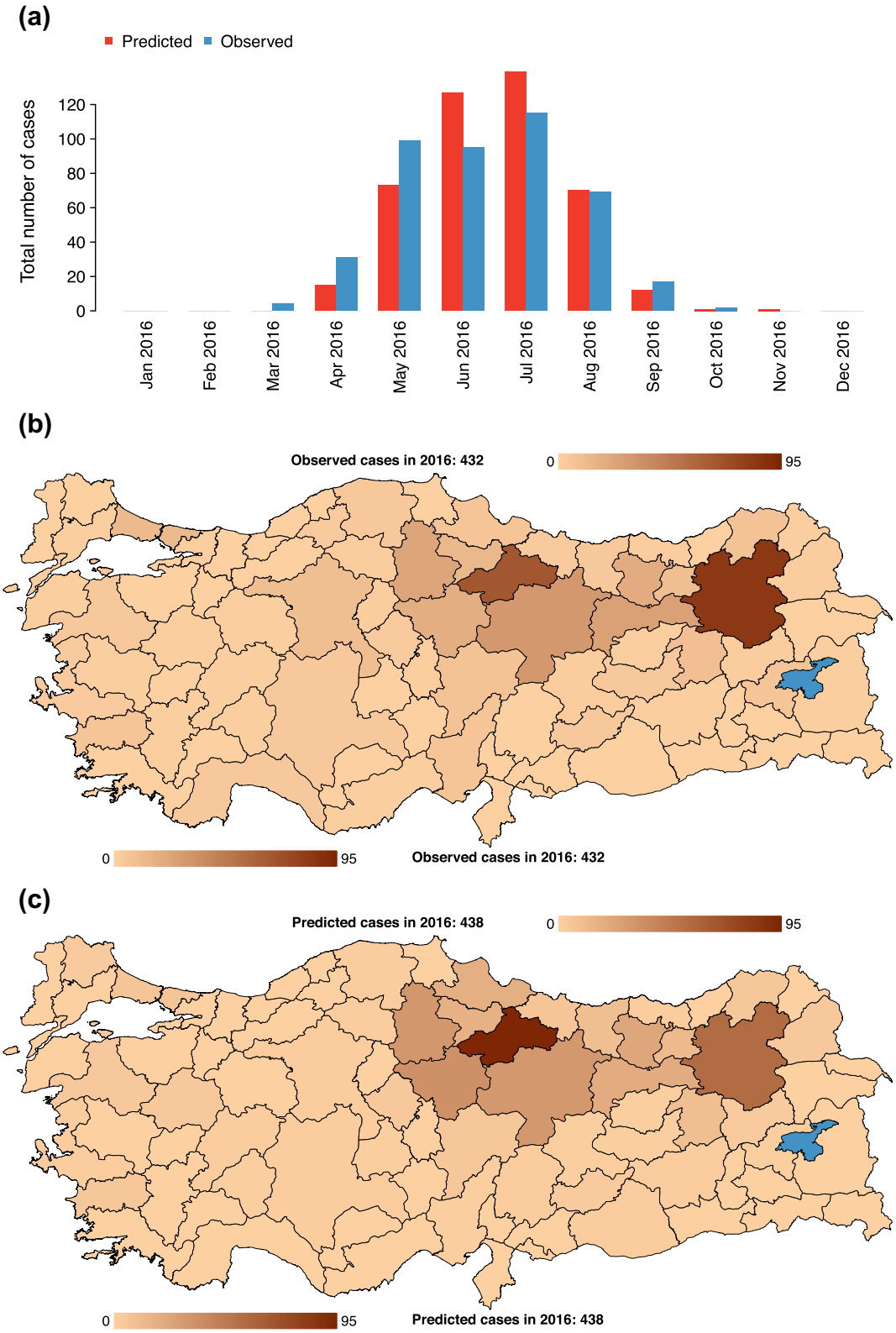


Fig. 2. Prediction results obtained by our structured Gaussian process algorithm for 2016. Observed cases are shown in blue and predicted cases are shown in red. (a) Monthly observed and predicted Crimean–Congo haemorrhagic fever (CCHF) case counts for 2016. (b) Annual observed CCHF case counts for each province in 2016. (c) Annual predicted CCHF case counts for each province in 2016. Numbers in the keys of (b) and (c) correspond to the minimum and maximum numbers of observed and predicted cases in provinces for 2016. Monthly prediction maps can be seen at http://midas.ku.edu.tr/ProspectiveCCHF/.

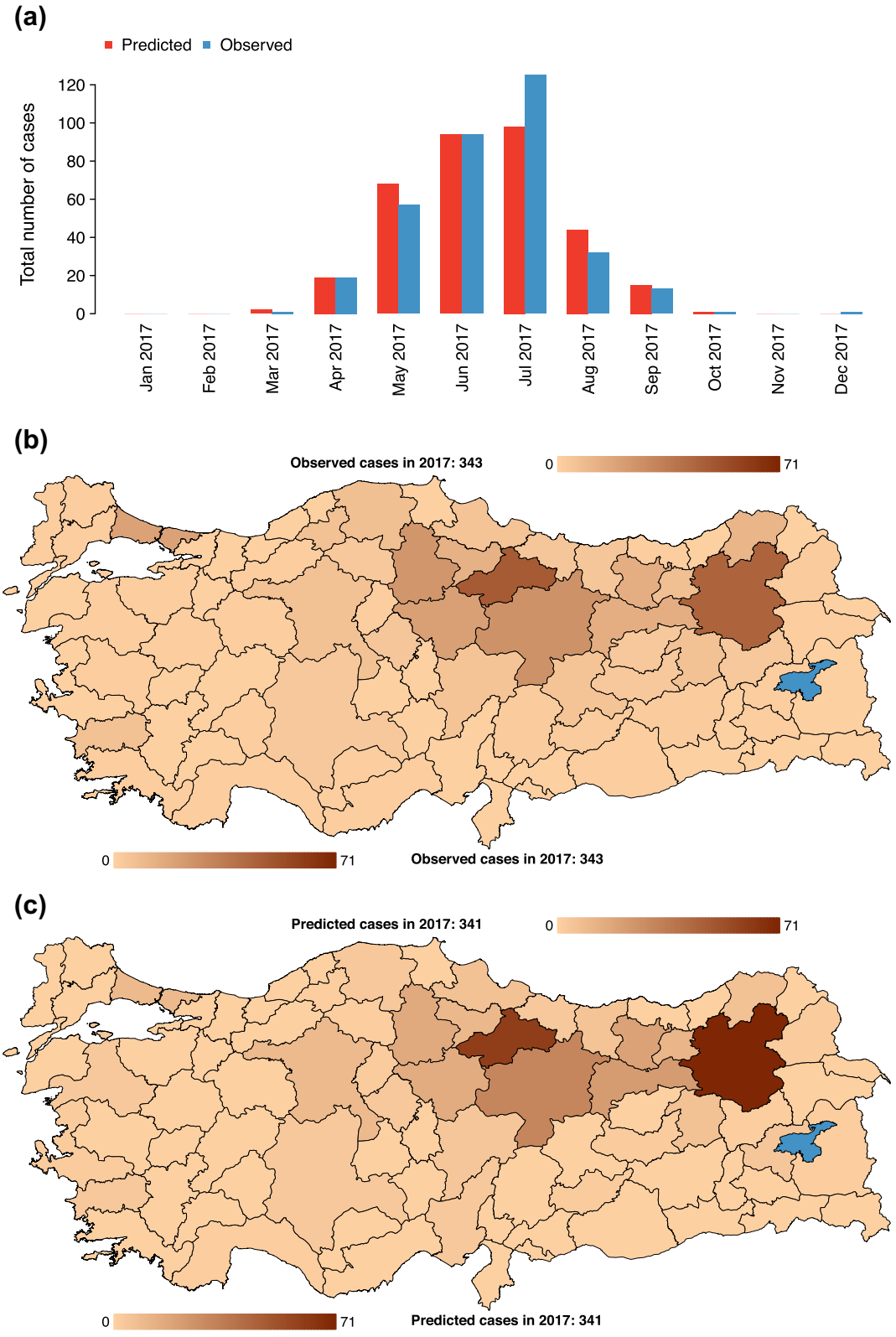


Fig. 3. Prediction results obtained by our structured Gaussian process algorithm for 2017. Observed cases are shown in blue and predicted cases are shown in red. (a) Monthly observed and predicted Crimean–Congo haemorrhagic fever (CCHF) case counts for 2017. (b) Annual observed CCHF case counts for each province in 2017. (c) Annual predicted CCHF case counts for each province in 2017. Numbers in the keys of (b) and (c) correspond to the minimum and maximum numbers of observed and predicted cases in provinces for 2017. Monthly prediction maps can be seen at http://midas.ku.edu.tr/ProspectiveCCHF/.

of spatial and temporal data sources to determine the relative importance of each data source. We evaluated our approach by performing monthly predictions for each province in a prospective manner.

The latitude, longitude and number of settlements with <25 000 inhabitants were found to be the most important spatial covariates for predicting CCHF case counts prospectively. Potential evapotranspiration and season were found to be the most informative temporal covariates for both the 2016 and the 2017 predictions.

The importance of number of settlements with <25 000 inhabitants could be related to the human population at risk living close to the habitat of ticks and animals as these settlements are usually situated in rural areas where people are engaged in agricultural activities. The number of settlements with <25 000 inhabitants is important for both years, but positions of latitude and longitude switched their rankings in terms of importance. This finding is in line with the increased number of CCHF cases in eastern parts in later years, which can be better captured by longitude rather than latitude.

Evapotranspiration is a climate variable and is defined as the total water vapour produced in the water basin as a result of the growth of plants in the water basin. Potential evapotranspiration is evapotranspiration at the time when there is sufficient water available to provide for a surface completely covered with plants. This term refers to providing the ideal amount of water to plants. It is also obvious that season covariate determines the temporal behaviour of CCHF or other seasonal infectious diseases in general. These two important temporal covariates confirm the role of the climate for the underlying mechanism of CCHF. Careful follow up of these covariates may provide possible warnings in the short term instead of having to wait for yearly predictions from our model. Higher temperature was previously found as a main driver for the abundance of *H. marginatum* [1,12,16,18] because high temperatures may accelerate the life cycle of ticks and so increase host questing.

In our study, we found that yearly changes in the land involving olive trees, fallow land and forest land were more important than the animal population (see Supplementary material, Fig. S1b,d). Our findings were parallel with those of another report in which the land cover, rather than climate and animal population, was found to be the main driver for world-wide distribution of CCHF. Those authors commented that these factors might be more important in predicting finer-scale prevalence patterns [12].

We used the annual data of husbandry from the Turkish Statistical Institute for the first time, and our model was able to reveal the importance of different animal groups (see Supplementary material, Fig. S1b,d). In our model for the 2017 prediction, goats, cattle and sheep were found to be the most significant animals for CCHF dynamics and spread, respectively. These findings contradict the observations of veterinarians in the field, who claim that bovine/cattle livestock are more important than goat livestock in the transmission cycle of the virus. This contradiction implies that there are some other underlying reasons such as the farmers; those caring for the goats might come into hand contact with them with or without protection. We must take into account the possible reasons why a covariate is chosen and take precaution against it respectively. The importance of covariates that may be related to human action indicates that awareness is lacking in some parts of the country about the presence of CCHF or precautions against CCHF. Our model identifies the directions to which we should pay close attention with high priority. For instance, in the areas with high goat, cattle or sheep density, agricultural workers and others working with animals should also be monitored and must be informed about CCHF. For further investigation, tick abundance studies in the field should be developed and improved.

Annual predictions for 2016 and 2017 are accurate, but the predictions for individual provinces are not as much accurate (Figs. 2 and 3). Predicting the total number of cases from overall seasonality is easier than capturing spatial dependencies because time-series data are dependent on whether there is seasonality behaviour of the data. One limitation of this study is that our model may not predict an outbreak if the reason for the outbreak is not related to the covariates that we used to train our model. However, when the first data of the outbreak arrive, the model will update itself accordingly, although there might be some delay for accurate predictions. Another limitation is that even if the surveillance data are ready, covariate data (e.g. livestock statistics) might be published much later or might be incomplete at the time of prediction. Then, the model would not be able to benefit from all information sources to better capture the progress of disease dynamics. However, these problems are valid for all data-driven models.

Our proof-of-concept study provided insights for understanding possible mechanisms of infectious diseases and found directions with high priority for practice and policy to combat against emerging infectious diseases. We tested our tool on a single disease, but the same framework can be extended towards other vector-borne infectious diseases, as well as other infectious diseases.

Transparency declaration

The authors declare no conflict of interests.

Funding

This work was funded by the Turkish Academy of Sciences (TÜBA-GEBiP; The Young Scientist Award Programme) and the Science Academy of Turkey (BAGEP; The Young Scientist Award Programme). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Access to data

The input covariates, nationwide CCHF surveillance data set and our computational results reported in this study are accessible at http://midas.ku.edu.tr/ProspectiveCCHF/.

Authors' contribution

ÇA, ÖE and MG designed the study and interpreted the results. ÇA collected and cleaned the spatial and temporal covariates used, implemented the software and generated the figures. ÇA and MG designed the software and performed the data analysis. ÇA and ÖE did the literature search, and drafted the first version of the paper, which was revised by MG. All authors contributed to the final version of the Article.

Acknowledgements

We are grateful to the Public Health Directorate of the Ministry of Health of Turkey for providing us with the nationwide CCHF surveillance data set.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cmi.2019.05.006.

References

[1] Ergonul O. Crimean–Congo haemorrhagic fever. Lancet Infect Dis 2006;6:203–14.
[2] Kilpatrick AM, Randolph SE. Drivers, dynamics, and control of emerging vector-borne zoonotic diseases. Lancet 2012;380:1946–55.
[3] Mostafavi E, Haghdoost A, Khakifirouz S, Chinikar S. Spatial analysis of Crimean–Congo hemorrhagic fever in Iran. Am J Trop Med Hyg 2013;89:1135–41.
[4] Ansari H, Shahbaz B, Izadi S, Zeinali M, Tabatabaee SM, Mahmoodi M, et al. Crimean–Congo hemorrhagic fever and its relationship with climate factors in southeast Iran: a 13-year experience. J Infect Dev Countries 2014;8:749–57.
[5] Vescio FM, Busani L, Mughini-Gras L, Khoury C, Avellis L, Taseva E, et al. Environmental correlates of Crimean–Congo haemorrhagic fever incidence in Bulgaria. BMC Public Health 2012;12:1116.
[6] Estrada-Pena A, Vatansever Z, Gargili A, Buzgan T. An early warning system for Crimean–Congo haemorrhagic fever seasonality in Turkey based on remote sensing technology. Geospat Health 2007;2:127–35.
[7] Estrada-Pena A, Zatansever Z, Gargili A, Aktas M, Uzun R, Ergonul O, et al. Modeling the spatial distribution of Crimean–Congo hemorrhagic fever outbreaks in Turkey. Vector Borne Zoonotic Dis 2007;7:667–78.
[8] Holmes EC, Rambaut A, Andersen KG. Pandemics: spend on surveillance, not prediction. Nature 2018;558:180–2.
[9] Washington ML, Meltzer ML, Centers for Disease C, Prevention. Effectiveness of Ebola treatment units and community care centers—Liberia, September 23–October 31, 2014. Morb Mortal Wkly Rep 2015;64:67–9.
[10] Bogoch II, Creatore MI, Cetron MS, Brownstein JS, Pesik N, Miniota J, et al. Assessment of the potential for international dissemination of Ebola virus via commercial air travel during the 2014 West African outbreak. Lancet 2015;385:29–35.
[11] Camacho A, Eggo RM, Goeyvaerts N, Vandebosch A, Mogg R, Funk S, et al. Real-time dynamic modelling for the design of a cluster-randomized phase 3 Ebola vaccine trial in Sierra Leone. Vaccine 2017;35:544–51.
[12] Messina JP, Pigott DM, Golding N, Duda KA, Brownstein JS, Weiss DJ, et al. The global distribution of Crimean–Congo hemorrhagic fever. Trans R Soc Trop Med Hyg 2015;109:503–13.
[13] Ç Ak, Ergönül Ö, Şencan İ, Torunoğlu MA, Gönen M. Spatiotemporal prediction of infectious diseases using structured Gaussian processes with application to Crimean–Congo hemorrhagic fever. Plos Negl Trop Dis 2018;12:e0006737.
[14] Harris I, Jones PD, Osborn TJ, Lister DH. Updated high-resolution grids of monthly climatic observations—the CRU TS3.10 Dataset. Int J Climatol 2014;34:623–42.
[15] Rasmussen CE, Williams CKI. Gaussian processes for machine learning. Cambridge, MA: MIT Press; 2006.
[16] Estrada-Pena A, Vatansever Z, Gargili A, Ergonul O. The trend towards habitat fragmentation is the key factor driving the spread of Crimean–Congo haemorrhagic fever. Epidemiol Infect 2010;138:1194–203.
[17] R Core Team. R. A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2017.
[18] Estrada-Pena A, Ruiz-Fons F, Acevedo P, Gortazar C, de la Fuente J. Factors driving the circulation and possible expansion of Crimean–Congo haemorrhagic fever virus in the western Palearctic. J Appl Microbiol 2013;114:278–86.