DISEASES SPREAD PREDICTION IN TROPICAL AREAS BY MACHINE LEARNING METHODS ENSEMBLING AND SPATIAL ANALYSIS TECHNIQUES

A. A. Kolesnikov [1] *, P. M. Kikin [2], A. M. Portnov [3]

[1] Siberian State University of Geosystems and Technologies, Novosibirsk, Russian Federation - alexeykw@yandex.ru
[2] Peter the Great St.Petersburg Polytechnic University (SPbPU), St.Petersburg, Russian Federation - it-technologies@yandex.ru
[3] Moscow State University of Geodesy and Cartography,Moscow, Russian Federation - portnov@miigaik.ru

KEY WORDS: disease spread, machine learning, ensembling, LSTM

ABSTRACT:

Infection with tropical parasitic diseases has a great economic and social impact and is currently one of the most pressing health problem. These diseases, according to WHO, have a huge impact on the health of more than 40 million people worldwide and are the second leading cause of immunodeficiency. Developing countries may be providers of statistical data, but need help with forecasting and preventing epidemics. The number of infections is influenced by many factors - climatic, demographic, vegetation cover, land use, geomorphology. The purpose of the research is to investigate the space-time patterns, the relationship between diseases and environmental factors, assess the degree of influence of each of the factors, compare the quality of forecasting of individual techniques of geo-information analysis and machine learning and the way they are ensembled. Also we attempt to create a generalized mathematical model for predicting several types of diseases. The following resources were used as a data source: International Society for Infectious Diseases, Landsat, Sentinel. The paper concludes with the summary table containing the importance of individual climatic, social and spatial aspects affecting the incidence. The most effective predictions were given by a mathematical model based on a combination of spatial analysis techniques (MGWR) and neural networks based on the LSTM architecture.

1. INTRODUCTION

Dengue is an infectious disease that is transmitted from person to person through the mosquitoes Aedes aegypti and Aedes albopictus, which are the main vectors of the virus in various parts of the world. The World Health Organization (WHO) estimates that about 50-100 million cases of dengue are reported worldwide every year, and two fifths of the world's population is at risk of epidemic. Dengue or DHF / DSS has affected more than a hundred countries. Since 1950, more than 500,000 hospitalizations and about 70,000 child deaths have been reported; the incidence rate among children reaches 64 per 1000 population.

According to the analysis of the global spread of dengue virus, the number of infections per year is estimated at 390 million, of which almost 96 million are symptomatic. It is estimated that the number of dengue infections has increased dramatically over the past 50 years, which has led to a huge impact on human health worldwide. Distribution regions include countries in South-East Asia, Latin America, Africa, where dengue has been hyper endemic for decades and is a serious problem (Kuno, 2007, Xiao, et al., 2016, Shepard, et al., 2013, Ooi, Gubler, 2009, Halstead, 2006).

Early prediction of the fever spreading risks and its quantitative characteristics will allow to carry out preventive measures and actions in order to reduce the risks and potential losses. However, at the moment there are no universal models that would make it possible to effectively make such forecasts for various territories. There are separate studies (Anno, et al. 2014, Kiang, Soebiyanto, 2012, Naish, Tong, 2014, Chan, et al, 2011) describing forecast models for certain territories, but the impossibility of extrapolating the results of these works to other territories significantly reduces their usefulness (The Influence of Global Environmental Change on Infectious Disease Dynamics, 2014, Gubler, 2011).

Thus, the main purpose of made research was to study the possibility of using machine learning methods to create a universal predictive model capable of predicting quantitative indicators of the incidence of dengue fever for various territories. Another goal was to assess the impact of the initial data set features on the final result of disease prediction.

2. METHODS AND MATERIALS

To create forecasting model, the remote sensing, cartographic and statistical data related to the environment collected by the Center for Disease Control and Prevention was used (Dengue and Climate, 2019), National Oceanic and Atmospheric Administration in the US Department of Commerce (Center for Disease Control, Health Map, 2019), and Philippine Department of Health was used (Department of Health, 2019).

The key predicted feature was the number of cases of dengue fever within 1 week. This feature can also serve as an indirect indicator for approximate estimation of fever outbreak probability. For example, very small values may indicate a low outbreak probability. However, its exact assessment requires the creating of a separate forecasting model.

To estimate the quality assessment of the developing model was selected the accuracy assessment metric - the mean absolute error. For its calculation, we used the service drivendata.org,

* Corresponding author

which allows us to verify the results of the model prediction for 2 territories, using a test sample provided by the service in Puerto Rico (San Juan) and Peru (Iquitos).

Initial data used to create forecasting model included the following features:

- city abbreviation;
- date of measurement;
- Current climate indicators and their forecast according to NOAA's GHCN, PERSIANN, NOAA's NCEP:
- temperature;
- humidity;
- rainfall;
- NDVI values for territories adjacent to the city, calculated from the pixels of the satellite image (Haug, Ostermann, 2014, Peters, et al, 2002, Bottou, 2010).

The sample included 936 records for the city of San Juan and 520 records for the city of Iquitos.

A visual representation of climatic parameters set and temporal coverage for the two described cities is shown in Figures 1-4.

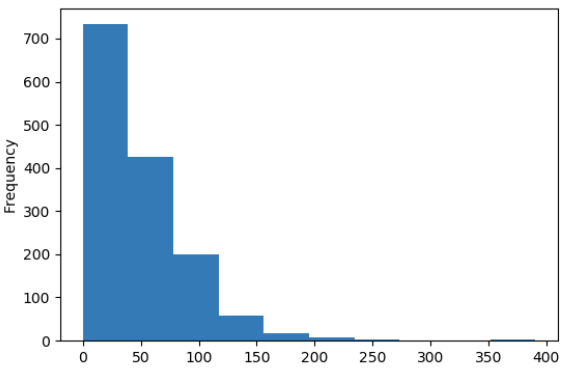

Figure 1. Average value of precipitation, mm

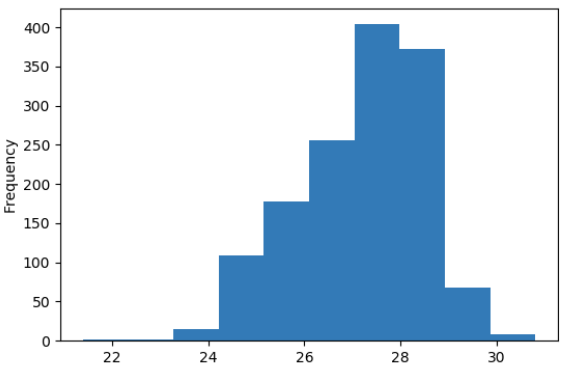

Figure 2. Average temperature, Celsius

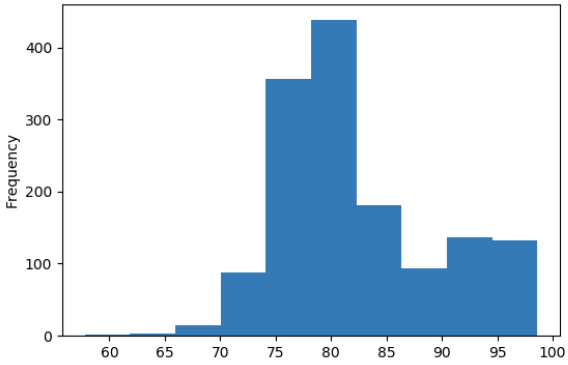

Figure 3. Relative humidity, percentages

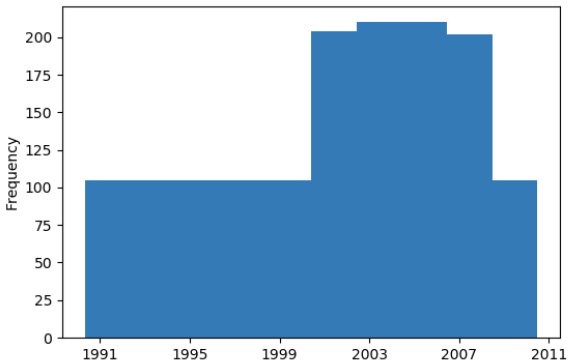

Figure 4. Dataset time interval, years

After forming the dataset, pre-processing of data was carried out, which included the removal of anomalous values - “outliers”, as well as normalization. The removal of anomalous values is necessary to eliminate from the sample, on the basis of which the predictive model is based, the data resulting from poor quality measurements. Such data can lead to a significant decrease in the accuracy of the forecast.

To detect the outliers Turkey’s fences method was used (Brown, 2002)**:**

- for each feature the quartiles were found (quartiles are the numbers which are ‘borders’ for separating the set of the features values to the 4 equal parts).
- the Interquartile range (IQR) (difference between 1 and 3 quartiles) was calculated.
- values which lies outside the extreme quartile plus 1.5 IQR are outliers.

Data normalization, i.e. bringing them to a single range of values is necessary for their use in metric algorithms that are sensitive to scaling, as well as in neural networks. Some algorithms, such as gradient boosting and random forest, do not require preliminary normalization of input data. In this regard, for each individual machine learning algorithm its own approach to data normalization was applied.

In order to select the most appropriate tools for solving the task, a fairly wide range of tools was analyzed - methods of geographic information systems, typical machine learning libraries, artificial neural networks.

In modern geographic information systems (ArcGIS Pro, GRASS GIS, SAGA GIS), such spatial forecasting methods as

IDW, local and global polynomial methods and kriging are used to simulate the spatial distribution of indicators of objects or phenomena. The disadvantage of these methods in terms of the task is the impossibility of taking into account the dynamics of changes in indicators over time while spatial prediction. In order to take into account these changes over time, one must either use specialized algorithms (for example, ARIMA / SARIMA), or convert information about the year, month, quarter, day, etc. in separate parameters with numeric values and add them to the attribute table as separate additional columns of the analyzed data. Using such data modifications, it is possible to take into account the dynamics of changes when using conventional regression analysis algorithms, for example, linear regression or random forest. Thus, for almost all geographic information systems, it is required either to connect additional modules (usually from Python or R), or to upload data for further processing. In this regard, the use of only GIS to solve this problem is not appropriate.

These technologies are well studied and are unspoken standards in the construction of analytical models, but if we talk about the prospects for development, then the newest direction for the problems of forecasting spatial-temporal data is neural networks. Neural networks have the ability to learn on heterogeneous data types that take into account both the spatial and temporal position of objects, which is of great theoretical and practical importance for creating models for analyzing and predicting time series. Additional advantages of neural networks are their high generalizing ability, which allows to work effectively in such non-standard conditions as non-obviousness of the internal data structure, errors and insufficiency in experimental data. Although neural networks are non-linear structures, they allow approximation of an arbitrary continuous function. A neural network based model can be trained in such a way that it could determines the further development of the process or phenomenon during the specified period with high confidence. Since the time series of most phenomena and processes are continuous functions, the use of neural networks in their prediction is fully justified and correct. The process of using neural networks is based on the use of tensorflow, theano, pytorch and some other libraries. In addition, there is the possibility of using the created models and scripts in GIS, since in many common geographic information systems the module development language is also python [19-23].

Features and data heterogeneity lead to the need to use machine learning methods. To solve the problem, we considered the following most popular machine learning methods: gradient boosting based on decision trees (in the implementation of xgBoost, LightGBM, CatBoost), random forest, nearest neighbor method, linear regression.

After a preliminary assessment of the data, the next step was the selection of parameters, so called – feature engineering (Russell, Norvig, 2010). This stage is very desirable for almost any dataset, since unnecessary features increase RAM requirements, reduce the model's learning rate and, most importantly, the ability to generalize, leading to model overfitting.

Feature engineering can be performed either by calculating numerical indicators of correlation and entropy, or visually, using charts, or a map. Comparison of parameters with each other can be performed using the traditional calculation of the correlation value (for example, Pearson and Spearman coefficients), as well as with the help of specialized calculated indicators focused on the analysis of spatial data and time series.

Spatial correlation is usually (Benedetti-Cecchi, et al, 2010, Bivand, et al, 2011, Fisher, Wang, 2011) measured using the Moran index, indicating whether there is a clustering of objects, or they are randomly arranged. The calculation of this indicator is implemented, for example, in ArcGIS Pro, GRASS GIS, PySAL.

For the analysis of the time series entropy, the Lyapunov index is the most universal. Also for this purpose the Hurst coefficient can be used, detrended fluctuation analysis (Song, et al, 2014, Li, et al, 2013, Kantz, Schreiber, 2004).

To assess the feature importance on the prediction result, a correlation matrix and conclusions on it were used based on the pandas profiling report. Also feature importance method (included in most implementations of scikit-learn algorithms) (Strobl, et al, 2008, Van der Laan, 2006) and a specialized algorithm for selecting parameters – Boruta were used (Kursa, Rudnicki, 2010).

According to the results of the correlation matrix, a high relation between the values of temperature, humidity and precipitation was revealed. (pairs of parameters:

- quarter - month ($\rho = 0.97069$),
- reanalysis_avg_temp_k - reanalysis_air_temp_k ($\rho = 0.90178$),
- reanalysis_sat_precip_amt_mm - precipitation_amt_mm ($\rho = 1$),
- reanalysis_specific_humidity_g_per_kg - reanalysis_dew_point_temp_k ($\rho = 0.99705$),
- reanalysis_tdtr_k - reanalysis_max_air_temp_k ($\rho = 0.91858$)).

According to the results of the importance assessment using Boruta and feature importances, the highest priority is given to temperature, NDVI index values and precipitation. The seven most important features according to the results of applying feature importance and Boruta methods are given in the table 1:

No.	Feature importance	Boruta
1	reanalysis_air_temp_k	ndvi_sw
2	ndvi_sw	ndvi_se
3	city	reanalysis_air_temp_k
4	reanalysis_dew_ point_temp_k	reanalysis_dew_ point_temp_k
5	reanalysis_tdtr_k	reanalysis_avg_temp_k
6	ndvi_se	reanalysis_tdtr_k
7	station_min_temp_c	station_min_temp_c

Table 1. Feature importances

The results of the feature analysis were then used in the construction and analysis of mathematical models based on Random Forest with the following criteria for selecting parameters: clipping parameters with a correlation of more than 0.9, clipping parameters with a Boruta rank of less than 20, clipping parameters with Feature importances <0.3. The results are shown in the final table.

At the stage of building forecasting models, the next most suitable algorithms were analyzed (Anselin, 2019, Davies, Van der Laan, 2016, Jiang, et al, 2017, Reichstein, et al, 2019):

- random forest;
- gradient booster based on decision trees (in the implementation of xgBoost, LightGBM, CatBoost);
- a neural network (in the implementation of Tensorflow and Keras) with two architectures distinguished by the presence of hidden layers (Goodfellow, et al, 2016);
- SARIMA and ensemble the results of the SARIMA algorithm;
- xgBoost.

During implementation of both neural networks preliminary data normalization was made. For the activation we used method "relu" and optimization method - "adam". Hidden layers of the neural network consisted of 5 and 13 elements (justification).

For the CatBoost algorithm, only the city was used as a categorical parameter and an additionaly year, season and week.

For a potential improvement of the prediction results, the original dataset was expanded with the distribution data (in numerical representation) (Kraemer, et al, 2015) of Aedes aegypti and Aedes albopictus over the world from 1958 to 2014. The results of the change in MAE, as a result of data expansion, are shown in the summary table.

Since one of the main objectives of the research was to assess the model scalability for different territories of its application regardless of the territory under consideration, so the best version of the constructed forecasting model was validated on the data of another territory. It contains the number of confirmed monthly cases of dengue fever in the municipality of Campinas (Spain) for the period from 1998 to 2015 (data source SES (Secretaria Estadual de Saúde) and SINAM (Sistema de Informaçao de Agravos de Notificação)), supplemented with information on precipitation in mm, average, maximum and minimum temperatures per day [https://www.kaggle.com/renangomes/dengue-temperatura-e-chuvas-em-campinassp].

3. RESULTS

The results of the constructed models quality assessment of are shown in Table 2.

Rating	Method and its application features	MAE
1.	LSTM and xgBoost Ensemble	25.1
2.	SARIMA and xgBoost Ensemble	25.8
3.	Random forest, separate models for each city, selection of hyperparameters, clipping parameters with a correlation of more than 0.9	26.38
4.	Random forest, separate models for each city, selection of hyperparameters, clipping parameters with a correlation of more than 0.9, additional data on mosquitoes	26.47
5.	Random forest, selection of hyperparameters, clipping parameters with a correlation of more than 0.9	26.5
6.	Random forest, selection of hyper parameters	26.6
7.	Random Forest in Orange	26.6130
8.	Random forest, selection of hyperparameters, clipping of attributes with a Boruta rank of less than 20	26.9
9.	Random forest, selection of hyperparameters, clipping of attributes with Feature importances <0.3	27.1
10.	нейронная сеть с двумя скрытыми слоями	27.4
11.	xgBoost, default options	27.9
12.	LightGBM, default options	28.7
13.	CatBoost, selection of hyperparameters, 5 categorical variables	29.561
14.	Linear regression in Orange	29.8173
15.	SARIMA	30.3
16.	Keras, no hidden layers	32.5
17.	KNN in Orange	33.8774
18.	Dataset Campinas (Spain)	36.7
19.	CatBoost, default parameters, 4 categorical features	37.1
20.	CatBoost, default parameters, categorical feature - city	37.2

Table 2. Models quality

4. CONCLUSION

According to the results of the research, the following conclusions were proposed:

- for spatial-temporal prediction with a large number of parameters, a combination of different algorithms for data processing using the methods of boosting or bagging is necessary;
- data preprocessing is usually no less complex and productive than building models;
- graphical tools for data processing and model building, such as Orange, are almost as good as python scripts, while exceeding their speed in creating models, but they have less capacity for data processing and presentation of results;
- random forest algorithms, along with gradient boosting, are the most universal for space-time forecasting tasks;
- in the case of a large number of inputs of the neural network, it is necessary to use hidden layers.

Also, in order to increase accuracy and model scalability, in further explorations it is planned to expand the initial data with open data on monitoring of mosquitoes Ae. aegypti and Ae. Albopictus, additional indices calculated on the basis of satellite monitoring. For building time-series based models it is going to use neural networks of the LSTM architecture and its variations.

REFERENCES

Anno, S., Imaoka, K., Tadono, T., Igarashi, T., Sivaganesh, S., Kannathasan, S., Kumaran, V., and Surendran, S.: Characterization of the Temporal and Spatial Dynamics of the Dengue Epidemic in Northern Sri Lanka, *Int. Arch. Photogramm. Remote Sens.* Spatial Inf. Sci., XL-8, 163-166, https://doi.org/10.5194/isprsarchives-XL-8-163-2014, 2014.

Benedetti-Cecchi L., Iken K., Konar B., Cruz-Motta J., Knowlton A, et al. Spatial relationships between polychaete assemblages and environmental variables over broad geographical scales. *PLoS ONE*, 2010, 5(9): e12946

Bivand R.S., Müller W., Reder M. Power calculations for global and local Moran's I. *Computational Statistics and Data Analysis*, 2009, 53: 2859-2872

Bottou L. Large-scale machine learning with stochastic gradient descent // *Proceedings of COMPSTAT' 2010*, Springer, 2010, pp. 177–186.
Brown, F.J., Reed C.B., Hayes J.M., Wilhite A.D., Hubbard K. A prototype drought monitoring system integrating climate and satellite data. *Proceedings of the Pecora L5/land satellite information 1V/ISPRS commission I/FIEOS*, 2002, Colarado, USA.

Center for Disease Control, Health Map (2019) DengueNet. www.healthmap.org/dengue/. (2019 Mar 15).

Davies M. M. and Van Der Laan M. J. Optimal Spatial Prediction Using Ensemble Machine Learning. *International Journal of Biostatistics*, 12(1):179–201, may 2016. ISSN 15574679. doi:10.1515/ijb-2014-0060. www.degruyter.com/view/j/ijb.2016.12.issue-1/ijb-2014-0060/ijb-2014-0060.xml

Dengue and Climate [Online]. Centers for Disease Control and Prevention, Communicable Diseases Center. www.cdc.gov/dengue/entomologyEcology/climate.html (30 April 2019).

Department of Health (DOH), www.doh.gov.ph. (10 Jul 2019).
E. H. Chan, V. Sahai, C. Conrad, and J. S. Brownstein, Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. *PLoS neglected tropical diseases*, vol. 5, no. 5, p. e1206, 2011.

Fischer M.M., Wang J.F. Spatial Data Analysis: Models, Methods and Techniques. 2010, New York: Springer
Goodfellow I., Bengio Y., Courville A. Deep Learning. MIT Press. 2016, 800 p. ISBN: 9780262035613

Gubler, DJ., 2011. Dengue, Urbanization and Globalization: The Unholy Trinity of the 21st Century, *Tropical Medicine and Health*, 39(4), Supplement, pp. 3-11.

Halstead S. Dengue in the Americas and Southeast Asia: do they differ? // *Revista panamericana de salud publica*. 2006. No. 20(6), pp.407–415.

Haug S., Ostermann J. A Crop Weed Field Image Dataset for the Evaluation of Computer Vision Based Precision Agriculture Tasks // *Computer Vision - ECCV 2014 Workshops.* Zurich: Springer, 2014, pp. 105–116.

Jiang Z., Li Y., Shekhar S., Rampi L., and Knight J. Spatial Ensemble Learning for Heterogeneous Geographic Data with Class Ambiguity. *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–10, New York, New York, USA, 2017. ACM Press. ISBN9781450354905. doi: 10.1145/3139958.3140044. dl.acm.org/citation.cfm?doid=3139958.3140044.

Kantz, H.; Schreiber, T. Nonlinear Time Series Analysis; Cambridge University Press: Cambridge, UK, 2004.

Kiang, R. K. and Soebiyanto, R. P.: Mapping the risks of malaria, dengue and influenza using satellite data, *Int. Arch. Photogramm. Remote Sens.* Spatial Inf. Sci., XXXIX-B8, 83-86, https://doi.org/10.5194/isprsarchives-XXXIX-B8-83-2012, 2012.

Kolesnikov A.A., Kikin P.M., Komissarova E.V., Grishenko D.V. Using machine learning for mapping. *Mezhdunarodnaya nauchno-prakticheskaya konferentsiya "Ot karty proshlogo -k karte budushchego"*, 28-30 Nov 2017, g. Perm' -g. Kudymkar. P. 110-120.

Kraemer Mu.G., Sinka M.E., Duda K.A., Mylne A., Shearer F.M., Brady O.J., Messina J.P., Barker C.M., Moore C.G., Carvalho R.G., Coelho G.E., Van Bortel W., Hendrickx G., Schaffner F., Wint Gr.W., Elyazar Ir.F., Teng H., Hay S.I. The global compendium of Aedes aegypti and Ae. albopictus occurrence. *Scientific Data*, *2017* 2(7): 150035. http://dx.doi.org/10.1038/sdata.2015.35

Kuno G. Research on dengue and dengue-like illness in East Asia and the Western Pacific during the First Half of the 20th century // *Reviews in medical virology.* 2007. No. 17(5):327–341. doi: 10.1002/rmv.545

Kursa, M. B. & Rudnicki, W. R. Feature selection with the Boruta package. *J STAT SOFTW* 36,1–13 (2010).

L. Anselin. A Local Indicator of Multivariate Spatial Association: Extending Geary's c. *Geographical Analysis*, 51(2):133–150, apr 2019. ISSN 15384632. doi: 10.1111/gean.12164. onlinelibrary.wiley.com/doi/abs/10.1111/gean.12164

Li, X.-J.; Hu, T.-S.; Guo, X.-N.; Zeng, X. Chaos analysis of runoff time series at different timescales. *J. Hydraul. Eng.* 2013, 44, 515–520.

Naish, S. and Tong, S.: Hot spot detection and spatio-temporal dynamics of dengue in Queensland, Australia, *Int. Arch. Photogramm. Remote Sens.* Spatial Inf. Sci., XL-8, 197-204, https://doi.org/10.5194/isprsarchives-XL-8-197-2014, 2014

Ooi E., Gubler D. Dengue in Southeast Asia: epidemiological characteristics and strategic challenges in disease prevention // *Cadernos de saude publica*. 2009;25 Suppl 1:S115–24.

Peters, J.A., Walter-Shea A.E., Ji L., Vina A., Hayes M., Svoboda D.M. Drought monitoring with NDVI-based standardized vegetation index. *Photogrammetric Engineering and Remote Sensing*, 2002. 68:7175.

Reichstein M., Camps-Valls G., Stevens B., Jung M., Denzler J., Carvalhais N., and Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743):195–204, feb2019. ISSN 14764687. doi: 10.1038/s41586-019-0912-1. www.nature.com/articles/s41586-019-0912-1

Russell, S. J.; Norvig, P. Artificial Intelligence: A Modern Approach (Third ed.). Prentice Hall. 2010. ISBN 9780136042594.

Shepard D.S., Undurraga E.A., Halasa Y.A. Economic and disease burden of dengue in Southeast Asia // *PLoS neglected tropical diseases.* 2013, No. 7(2):e2055 PubMed Central PMCID: PMCPmc3578748. doi: 10.1371/journal.pntd.0002055

Song, X., Zhang, Z., Chen, Y., Wang, P., Xiang, M., Shi, P., Tao, F. Spatiotemporal changes of global extreme temperature events (ETEs) since 1981 and the meteorological causes. *Nat. Hazards* 2014, 70, 975–994.

Strobl C., Boulesteix A., Kneib T., Augustin T., and Zeileis A. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):307, 2008.

The Influence of Global Environmental Change on Infectious Disease Dynamics: Workshop Summary. Washington (DC): National Academies Press (US); 2014 No.3. https://www.ncbi.nlm.nih.gov/books/NBK241611/ (11 June 2019).

Van der Laan M. J. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1):1008, 2006.

Xiao J.P., He J.F., Deng A.P., Lin H.L., Song T., Peng Z.Q., Characterizing a large outbreak of dengue fever in Guangdong Province, China. Infectious diseases of poverty. 2016. // *PubMed Central* PMCID: PMCPmc4853873. doi: 10.1186/s40249-016-0131-z

Revised July 2019