Handling missing data and validating models under high data scarcity for data analysis and modelling of climate drivers of vector borne diseases

Milica Tosic[3], Vasilije Matic[2], **Suzana Blesic**[1]

[1]Faculty Of Physics, University Of Belgrade, Belgrade, Serbia

We present results of our data analysis and modelling work to better understand and explain climate and environmental drivers of two vector-borne diseases or disease groups—of malaria, transmitted by mosquitoes, in Limpopo province of South Africa, and of sand fly-borne diseases in the Iberian Peninsula. Missing data and model validation challenges dominate vector-borne disease analytics both in these regions and more generally. Our approach draws on statistical-physics tools developed for long-range correlated and multifractal time series, applied here to the climate and health interface.

For the mosquito-borne case, we analyzed weekly malaria hospital admissions from the period 2000–2020, from five Limpopo districts, and cross-correlated them with temperature, precipitation, and evapotranspiration records. We performed wavelet transform spectral analysis (WTS), combined with a superposition-of-signals rule developed for long-range correlated data, to delineate characteristic time lags between meteorological drivers and case counts. We incorporated these lags into a regression model with multivariate climate drivers and a hazard-function term derived from long-range memory statistics to capture instances of extreme case numbers. The critical values we retrieved for temperature, rainfall, and bare-soil evaporation align with known biology of disease-carrying mosquitoes and their pathogens, providing initial validation. Finally, we generated projections for 2021–2050 and 2051–2080 under RCP2.6 and RCP8.5 climate scenarios to understand spatial expansion of climate suitability for malaria.

For the sand fly case, we applied the same analytical framework to Iberian weekly entomological monitoring and veterinary clinical records, combined with ERA5-Land climate variables and hydrological data from the mHM model. In this case, the principal challenge was data scarcity and heterogeneity: the sand fly monitoring (trapping) series and veterinary records that were used as inputs to our analysis differ greatly in sampling frequency, spatial coverage, and recording protocols, with many time series spanning only several months to a few years, insufficient for detection of long-term trends, seasonal cycles, or rare events. To circumvent this, substantial preprocessing, cross-source harmonization, and integration of auxiliary environmental datasets are required as preconditions for any attempt at modelling. We will present some results of these efforts and discuss how the superposition-rule and hazard-function methodology developed for the malaria case can be adapted to short, heterogeneous records.