Noise2NAKO: AI Methods Linking Environment and Health - a Large-Scale Cohort Application

Sahar Behzadi[1], Kathrin Wolf[1], Annette Peters[2], Alexandra Schneider[1], Mahyar Valizadeh[3], Wolfgang zu Castell[3], Jeroen Staab[4], Hannes Taubenböck[4]

[1] Institute of Epidemiology - Research Group „Environmental Risk“, Helmholtz Zentrum München, Neuherberg, Germany; [2] Institute of Epidemiology, Helmholtz Zentrum München, Neuherberg, Germany; [3] Department of Information and Communication Technology (ICT), Helmholtz Zentrum München, Neuherberg, Germany; [4] German Aerospace Center (DLR), Munich, Germany

The environment has major impacts on human health which requires sophisticated models to better reflect life exposures and to determine the long-term impacts of environmental factors on health. Thus, advanced statistical and data science approaches are needed to understand the complex interplay between the environment and population health. Existing models are hampered by the trade-off between complexity, interpretability and the biased nature of population-based data. Studies reported adverse impacts of noise on human health, mainly for cardiovascular outcomes, e.g. hypertension. Several studies also found a systematic bias w.r.t. social status. In this project, we aim at developing beyond state-of-the-art ML methods to advance existing noise maps, improve the quantification of noise impacts on health and delineate the complex interplay between environmental, socio-economic and health data. As a basis, we compile extensive German-wide noise maps applying data augmentation and deep CNN to overcome the spatial limitations of existing maps. Next, the noise data will be linked to socio-economic and demographic data from participants of the German national cohort (NAKO). In a first prediction task, we will identify German-wide vulnerable clusters in terms of noise and neighborhood factors for the risk of hypertension employing Distribution Regression Networks. In a second task, we will enhance this network by individual socio-economic and health data to investigate the interplay of different risk factors on hypertension. That is, we test and expand interpretable ML techniques e.g. AdaBoost, random forest and QRF to our setting and compare them to traditional models, e.g. additive logistic regression. To enhance the quality of data, we consider pre-processing methods, e.g. standardization, normalization and imputing missing values. Moreover, we will employ effective feature selection approaches, e.g. correlated features or information theoretic-based algorithms.

Keywords: Ensemble Learning, Deep Learning, Distribution Regression Network, Regression, Deep Regression Model, Deep Neural Network, Ensemble Methods, Interpretable AI/ML, Earth Observation, Satellite, Noise, Health, NAKO Cohort, Hypertension