TU Dortmund, Fakultät Statistik, Lehrstuhl Statistik in den Biowissenschaften
Robert Koch-Institut, Berlin

# Abschlussarbeiten im Bereich „Zeitreihenanalyse in der Epidemiologie”

Das Robert Koch- Institut (RKI) in Berlin (Dr. Stéphane Ghozzi, Dr. Alexander Ullrich) vergibt in Zusammenarbeit mit dem Lehrstuhl für Statistik in den Biowissenschaften (Prof. Dr. Roland Fried) mehrere Themen für Abschlussarbeiten (bevorzugt als Masterarbeiten). Nachfolgend finden sich Kurzbeschreibungen mehrerer Themenvorschläge. Die Betreuung erfolgt gemeinsam durch beide Einrichtungen. Hierfür sind auch längere Aufenthalte in Berlin vorgesehen, typischerweise zwischen 3 und 6 Monaten (länger auch möglich). Eine kleine finanzielle Unterstützung von bis zu 800 € monatlich (entspricht den maximal erlaubten monatlichen 80 Stunden) ist gewährleistet.

Weitere Auskünfte erteilen
Roland Fried fried@statistik.tu-dortmund.de
Stéphane Ghozzi GhozziS@rki.de https://de.linkedin.com/in/stephaneghozzi
Alexander Ullrich https://www.researchgate.net/profile/Alexander_Ullrich
Siehe auch www.rki.de/signale-project

## 1. Projection, clustering and labelling of individual cases

The RKI has for each reported infection case a lot of informations, such as age, sex, symptoms, lab results and more. This information is high dimensional and heterogeneous. One usually aggregates cases according to one or two dimensions (typically week, place, age, sex) and represents the case counts in time series, maps or histograms. Analyses are usually done on univariate time series.

But epidemiologists always consider individual cases in the end. Can one find an appropriate representation and develop corresponding case-based analyses? The idea would be to start and represent cases as a cloud of points in two dimensions where similar cases are close to one another. Possible approaches for this are based on dimension-reduction algorithms such as t-SNE or UMAP. One has to define an epidemiologically relevant distance measure.

Then cases that belong to a known outbreak can be labeled. This allows to visualize outbreaks and cases in a simple way and formulate hypotheses whether certain cases, that have not been assigned to an outbreak but are close to cases that have been, don't also belong to an outbreak.

In a second step, a classification algorithm should be developed, implemented and evaluated, that assigns to each case a probability to belong to a given outbreak. This could greatly help outbreak detection and investigation. The results (dimension reduction and labelling) should be presented in an interactive widget that will be integrated in an epidemiological web application. Ideally, the user will be able to set herself the parameters of the representation and the analysis.

References:

1. https://distill.pub/2016/misread-tsne/ (Tips on t-SNE usage)
2. https://lvdmaaten.github.io/tsne/ (t-SNE implementations)
3. https://arxiv.org/abs/1802.03426 (UMAP theory)
4. https://github.com/lmcinnes/umap (UMAP Practical explanation and implementation)
5. https://calculatedcontent.com/2012/10/09/spectral-clustering/ (Spectral clustering explanation)
6. https://en.wikipedia.org/wiki/Clustering_high-dimensional_data (curse of dimensionality, overview of clustering approaches)

## 2. Classification of diagnoses

Through the AKTIN project the RKI has access to emergency-department data: Information on the patients, the reason they came, quantitative measurements such as body temperature, and more. The diagnoses encoded in an ICD code is known for those that have been hospitalized only, or about a third of the patients.

It would be useful to learn the missing diagnoses: To gain insights in the determinants of an inflection (and maybe assist physicians and the hospital, suggesting diagnoses and which services might have patients coming) and to enrich the data set. In particular this could help compare it with other data sets, such as the case count of notifiable diseases.

A first simple study on a small data set has shown that influenza codes can be learned with a passable performance using logistic regression and that the enriched data set better correlates with reported influenza cases. This will be investigated further by using a series of classical classification algorithms and systematically train, test and compare them. They will be applied to different groups of codes. The time-series analysis comparing both data sets (AKTIN and notifiable diseases) will be refined. This approach could be then applied to other data sets at the RKI.

References:

1. http://win.ua.ac.be/~adrem/bibrem/pubs/confcov.pdf (prediction of ICD-10 codes)
2. Are Mortality and Acute Morbidity in Patients Presenting With Nonspecific Complaints Predictable Using Routine Variables? M.A. jenny et al., 2015, Academic Emergency Medicine, https://doi.org/10.1111/acem.12755

## 3. Signal detection with GLMs: implementation, performance and communication of models that incorporate time-space dynamics, duration (cumulative count), reporting delays

The automatic outbreak detection at the RKI is at it's a core a simple aberration detection on univariate time-series. This ignores a number of important features of infection dynamics. Theoretical approaches have been suggested to integrate these that remain in the simple framework of GLMs. The different models proposed will be implemented, augmented to produce signals when an unexpectedly high number of cases is reported, combined and evaluated.

Part of this project will also be the efficient communication through visualisations of these new features: space and time extension of a possible outbreak, its origin, how to best rank signals, etc.

References:

1. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2818 (multivariate SATSCAN)
2. https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020059 (space-time SATSCAN)
3. https://arxiv.org/pdf/1411.0416.pdf (review of spatio-temporal regression models and point processes)
4. https://academic.oup.com/biostatistics/article/18/2/338/2739329 (multivariate regression model)

### 4. Forecast and scenarios, i.e. event-conditioned or parameter-adjustable forecasts (GLM, SIR, random walk, HTM, agents, . . . )

Develop and implement parameter-adjustable forecasts of infectious-disease dynamics and communicate scenarios. Different methods will be applied and compared and an interactive widget developed to communicate the results.

This will include gathering requirements from the users (epidemiologists): Which methods do they favor, which time horizon is relevant for the forecast, which kind of event should be modeled, etc.

References:

1. https://en.wikipedia.org/wiki/Scoring_rule#Proper_scoring_rules
2. https://www.sciencedirect.com/science/article/pii/S0169207018300785 (insights from forecasting competition)
3. https://delphi.midas.cs.cmu.edu/~dfarrow/thesis.pdf (thesis by the 3-time influenza forecast winner)
4. http://reichlab.io/blog/ (practical discussion of influenza forecasting)
5. https://otexts.org/fpp2/ (forecasting textbook with R examples)

### 5. User feedback: recommender system and reinforcement learning

We are currently developing an epidemiological web application that will include dashboards communicating possible outbreaks ("signals") but also different analysis results and data visualisations. This application will be personalisable: a user will have an account and be able to set preferences.

We want to deepen the interaction with the user in two ways. First, a recommender system should make suggestions about what might be interesting at a given time (based on own past behaviour or the behaviour of similar users). One might also consider suggesting literature or press articles. Second, we want to collect feedback from the users, especially about whether the signals detected are relevant. This could be used then to correct the signal detection algorithms (reinforcement learning).

One or both systems will be developed prototypically and workflows will be defined. If the epi web app is already mature enough, one might think about implementing them.

References:

1. https://hackernoon.com/introduction-to-recommender-system-part-1-collaborative-filtering-singular-value-decomposition-44c9659c5e75 (collaborative filtering explanation)
2. https://www.ncbi.nlm.nih.gov/pubmed/27318069 (prediction market in disease forecasting)
3. https://delphi.midas.cs.cmu.edu/~dfarrow/thesis.pdf (thesis on influenza forecasting with wisdom of the crowd method)
4. https://unanimous.ai/wp-content/uploads/2018/09/ASI-for-Radiology-IEEE-IEMCON-2018.pdf (hybrid system AI-specialists)