Master thesis summary: Spatio-temporal forecasts for bike availability in dockless bike sharing systems

The goal of my master thesis was to develop a spatio-temporal forecasting model for bike availability in dockless bike sharing systems (DBSS). Forecasting where and when bikes will be available can assist system operators in anticipating on inbalances between supply and demand, and at the same time give users the option to plan their trips in advance. As such, forecasting bike availability is important to turn the shared bike into a reliable mode of transport. At the time I wrote my thesis, early 2019, the new generation of dockless bike sharing systems just appeared, and research on the topic was still very scarce.

In DBSS, the bikes can be picked up and dropped off at any location within a service area and any time within the service hours. This makes the forecasting task inherently spatio-temporal in nature. Where in conventional station-based systems forecasts are only needed at fixed locations in space (i.e. they usually only consider the time dimension as being variable), DBSS require a forecasting model that can provide a forecast at *any* point in space. This could be addressed by using a grid-based approach in which the spatial dimension is discretized, and the model forecasts the number of available bikes inside a grid cell at a given timestamp in the future. Instead, I opted for a distance-based approach. In this approach, the model forecasts the expected distance to the nearest available bike, for a given point in space at a given timestamp in the future. This approach does not suffer from the modifiable areal unit problem nor from edge effects, and allows users to define for themselves if the distance is acceptable or not. Furthermore, it forecasts continuous variables instead of counts, allowing for a wider choice of statistical forecasting techniques.

For the forecasting task I used Auto-Regressive Integrated Moving Average (ARIMA) models, which have their origin in the field of statistical time series analysis. Applying such models requires to estimate several parameters by fitting the model to a observed time series extracted from historical data. Model fitting can be a time consuming process, especially when working with long and dense time series. In my thesis, I tried to find a balance between the accuracy of the forecast and the speed at which they can be produced. This lead to a step-wise workflow in which model parameters are pre-fitted. In the first step, historical time series of distances to the nearest available bike are extracted for a dense set of points in space. These points are spatially clustered based on the similarity of the patterns in their time series. Then, a single ARIMA model is fitted for each spatial cluster. Whenever a forecast is requested for a specific location in space and time, the model parameters are taken from the pre-fitted model belonging to the cluster in which this location lies. Model parameters are updated regularly, but don't have to be fitted at each individual forecast request.

I tested the proposed forecasting workflow with historical data from a DBSS in San Francisco. The results showed that the model clearly outperformed simpler baseline methods and was especially performant in areas where the usage of the system was high. However, they also emphasized the limited forecastability of DBSS in general. This is mainly due to the irregular usage patterns of DBSS when compared to station-based systems. The higher the entropy of the time series, the harder it becomes to accurately forecast them. Suggestions where made to improve the accuracy of the system, for example by including the influence of exogenous variables such as weather conditions.

The thesis was supervised by Edzer Pebesma, Professor of Geoinformatics at the University of Münster. It was co-supervised by Jorge Mateu, Professor of Statistics at University Jaume I in Castellón, and Joel Silva, post-doctoral researcher at the New University of Lisbon.

The thesis document was written in a reproducible way with RMarkdown. You can find it on GitHub: https://github.com/luukvdmeer/msc-thesis. All the R code used for calculations is bundled in an open-source R package which can be found on GitHub as well: https://github.com/luukvdmeer/dockless.