Multi-Output Gaussian Processes for Crowdsourced Traffic Data Imputation

Traffic speed data imputation is a fundamental challenge for data-driven transport analysis. In recent years, with the ubiquity of GPS-enabled devices and the widespread use of crowdsourcing alternatives for the collection of traffic data, transportation professionals increasingly look to such user-generated data for a good deal of analysis, planning, and decision support applications. However, due to the mechanics of the data collection process, crowdsourced traffic data such as probe-vehicle data is highly prone to missing observations, making accurate imputation crucial for the success of any application that makes use of that type of data. In this paper, we propose the use of multi-output Gaussian processes (GPs) to model the complex spatial and temporal patterns in crowdsourced traffic data. While the Bayesian nonparametric formalism of GPs allows us to model observation uncertainty, the multi-output extension based on convolution processes effectively enables us to capture complex spatial dependencies between nearby road segments. Using six months of crowdsourced traffic speed data or "probe vehicle data" for several locations in Copenhagen, the proposed approach is empirically shown to significantly outperform popular state-of-the-art imputation methods.