Spatial-Temporal K Nearest Neighbors Model on MapReduce for Traffic Flow Prediction

A. Agafonov, A. Yumaganov

Samara National Research University

The 19th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2018)

Task definition

- Forecast the traffic flow in 10 minutes ahead
- Take into account spatial and temporal characteristics of the traffic flow
- Develop a distributed forecasting model
- Efficiently process large-scale traffic data

Task

- Real-time processing
- High accuracy

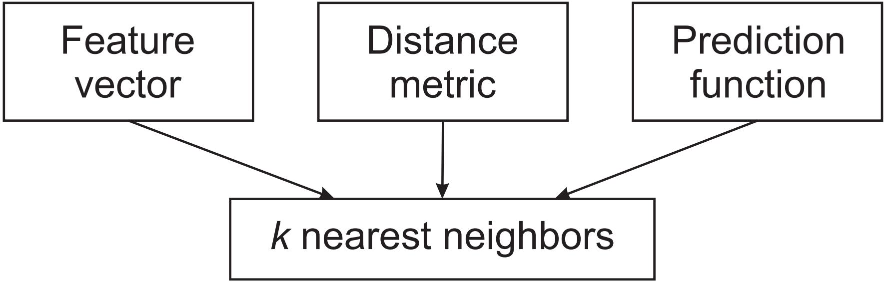
Problem formulation

- $G = (N, E)$ is a directed graph representing the road network;
- N is a node representing the road intersection;
- E is an edge denoting the road segment;
- V_t^j is an observed traffic flow characteristic on an edge $j \in E$ in a time moment t.

Given a graph $G(N, E)$ and traffic flow data $V_t^j, j \in E, t = 1, 2, \ldots T$, predict the traffic flow characteristic at a time interval $(t + \Delta)$ for a predefined prediction horizon Δ.

Proposed model

A short-term traffic flow forecasting model based on non-parametric regression *k* nearest neighbors algorithm is proposed.



Feature vector

Time-Domain Upstream / Downstream (TDUD) feature-vector:

$$(V^j_{t-T}, \ldots, V^j_{t-1}, V^j_t, V^{j-1}_{t-T}, \ldots, V^{j-1}_{t-1}, V^{j-1}_t V^{j+1}_{t-T}, \ldots, V^{j+1}_{t-1}, V^{j+1}_t)$$

Proposed feature vector:

- Partition the transportation network graph into several spatially compact clusters $\{G_i\}$ and define the cluster feature vector

$$\{V^j_t\}, j \in G_i, t = t_{cur} - T, \ldots, t_{cur}$$

- Reduce the dimensionality of the cluster feature vector using PCA procedure

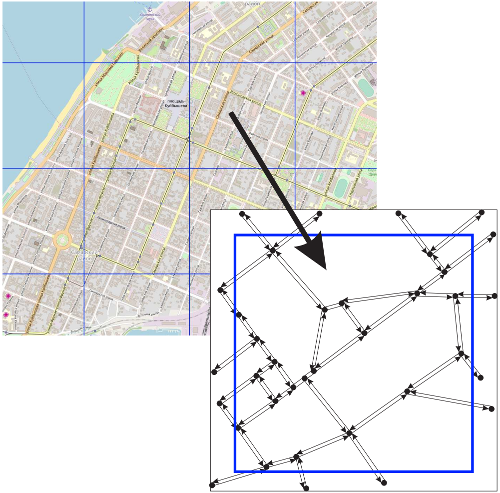
$$\{X_n\}^i, n = 1, \ldots, N$$

- Define the result feature vector for each road segment $j \in E$

$$S_j = (\{V^j_t\}, \{X_n\}^i), \quad i : j \in G_i, \quad t = t_{cur} - T, \ldots, t_{cur}, \quad n = 1, \ldots, N.$$

Graph partitioning

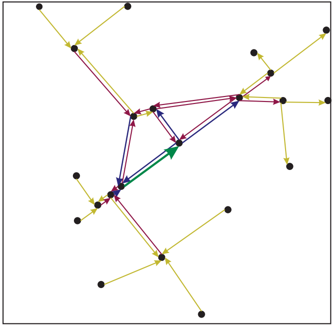
Partitioning by area G^{area}



Partitioning by distance G^{dist}

$$G_i^{dist} = \{j \in E : r(i,j) <= R\},$$

where $r(i,j)$ is the distance, $i \in E, j \in E$



Proximity measure

Weighted Euclidean distance with the trend adjustment:

$$d(S, \bar{S}^i) = d^{link}(V, \bar{V}^i) + \gamma d^{pca}(X, \bar{X}^i),$$

$$d^{link}(V, \bar{V}^i) = \alpha \sqrt{\sum_{t=1}^{T} \beta^{T-t+1} \left(V_t - \bar{V}_t^i\right)^2} + (1-\alpha) \sqrt{\sum_{t=2}^{T} \sum_{\delta=1}^{t-1} \left((V_t - V_\delta) - \left(\bar{V}_t^i - \bar{V}_\delta^i\right)\right)^2},$$

$$d^{pca}(X, \bar{X}^i) = \sqrt{\sum_{n=1}^{N} \left(X_n - \bar{X}_n^i\right)^2}.$$

Prediction function

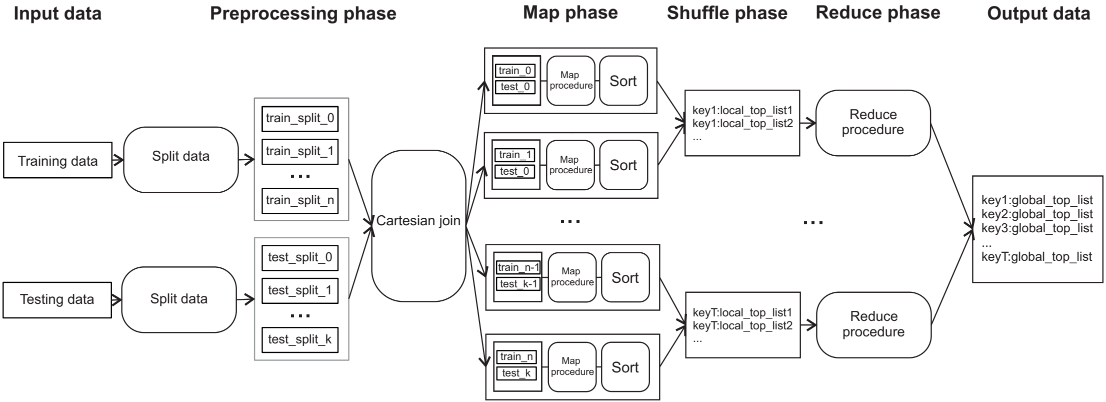
Prediction function by the weighted average:

$$\hat{V}_{T+1} = \sum_{k=1}^{K} \frac{d_k^{-1}}{\sum_{k=1}^{K} d_k^{-1}} V_{T+1}^{k}$$

Prediction function that combines the weighted average and the trend adjustment:

$$\hat{V}_{T+1} = \partial \sum_{k=1}^{K} \frac{d_k^{-1}}{\sum_{k=1}^{K} d_k^{-1}} V_{T+1}^{k} + (1 - \partial)\left(V_T + \frac{1}{KT} \sum_{k=1}^{K} \sum_{t=1}^{T} \left(V_{T+1}^{k} - V_t^{k}\right)\right)$$

MapReduce-based implementation



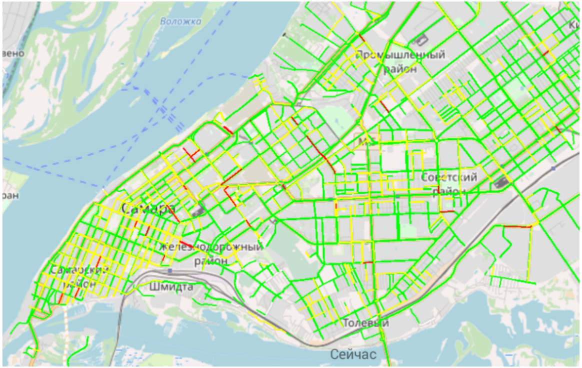
Model analysis

Comparison:

- proposed kNN model
- TDUD-KNN
- SARIMA

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|V_t - \hat{V}_t|,$$

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|V_t - \hat{V}_t|}{V_t} \times 100\%$$



Data set:

- Transportation network with 26018 road segments
- Average speed in a period of 60 days
- New data each 10 minutes

Model analysis. MAE / MAPE

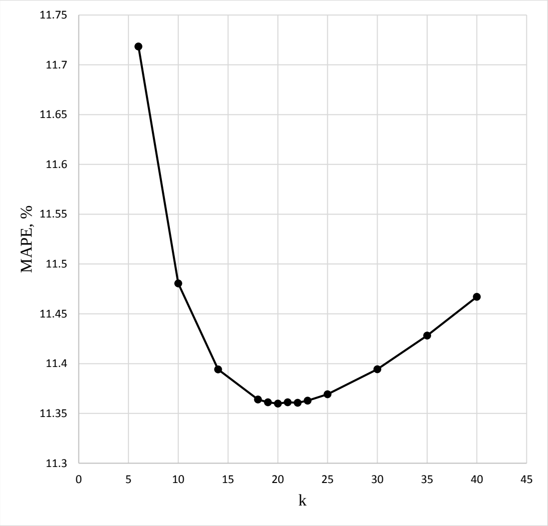
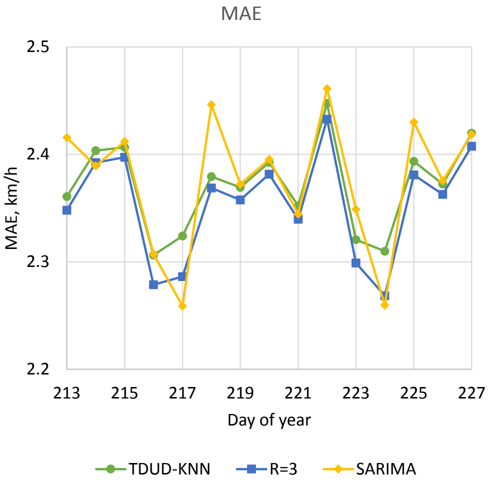
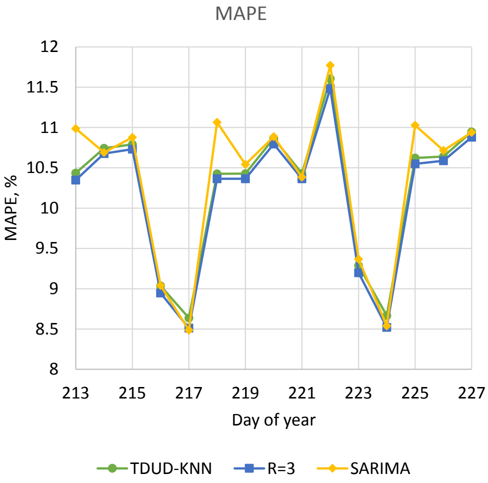


Table: Algorithms Comparison

	MAE	MAPE
$R = 1$	2.378	10.61
$R = 2$	2.374	10.598
$R = 3$	**2.372**	**10.593**
G^{area}	2.379	10.596
TDUD-KNN	2.387	10.611
SARIMA	2.399	10.77

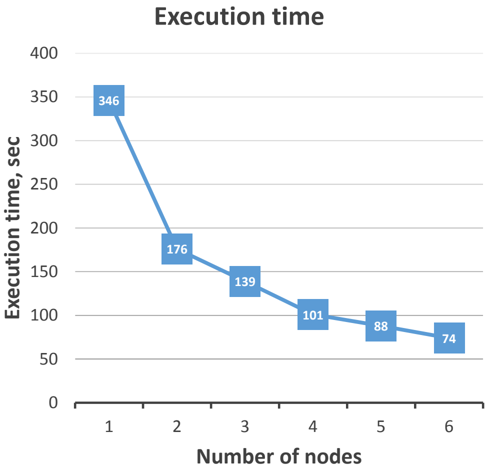
Model analysis. MAE / MAPE by days

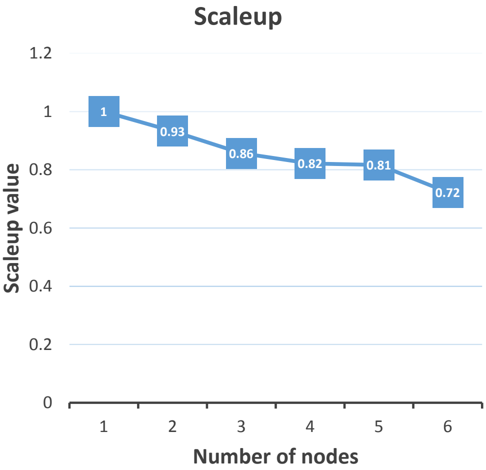




Model analysis. Execution time

Cluster up to 6 PC: Intel Core i5-3740 3.20 GHz, 8 GB RAM





Conclusion

The distributed spatial-temporal model of short-term traffic flow forecasting has the following advantages:

- The model takes into account spatial and temporal characteristics of the traffic flow.
- The implementation is based on MapReduce processing model in the open-source cluster-computing framework Apache Spark for distributed Big Data processing.
- The proposed model has a high prediction accuracy and reasonable execution time, sufficient for real-time prediction.

Thank you!

Anton Agafonov
ant.agafonov@gmail.com

The work was supported by the Ministry of Science and Higher Education of the Russian Federation (project no. RFMEFI57518X0177)