Методы оперативной классификации в задаче анализа данных дорожного движения

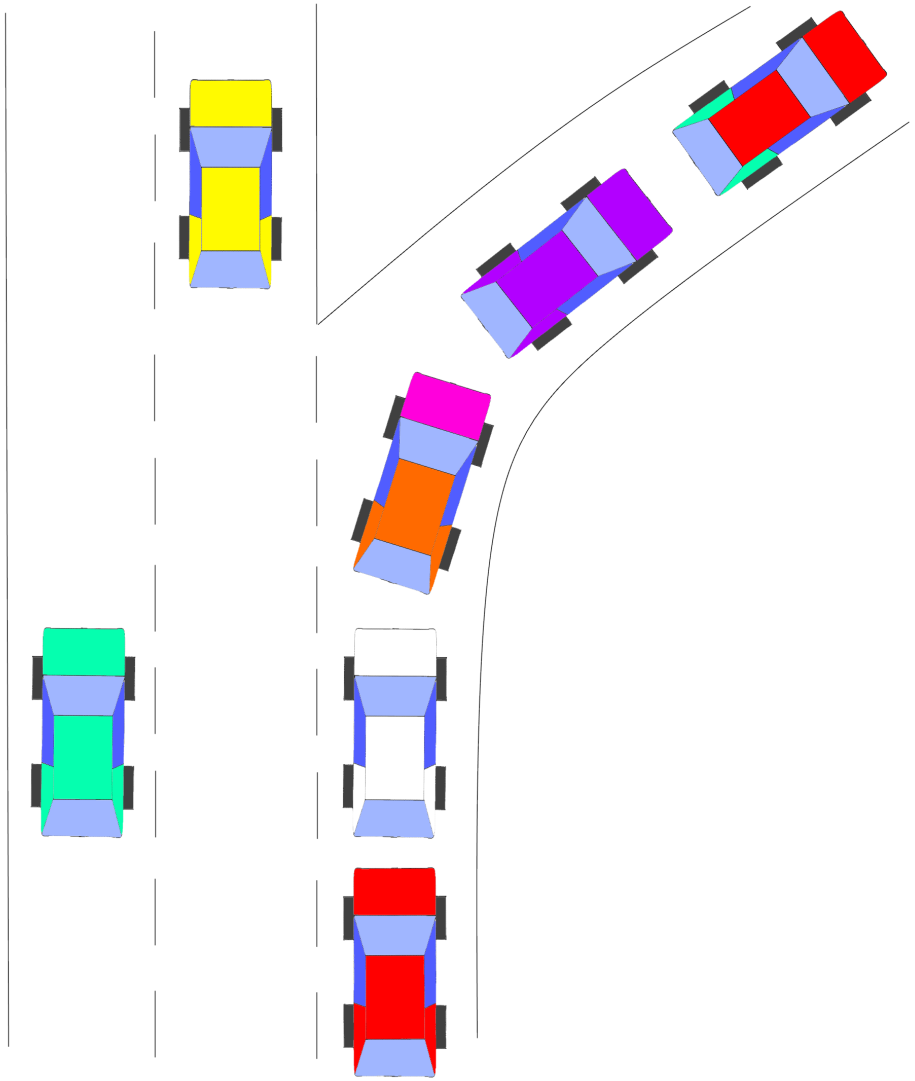
Игорь Олегович Молибог

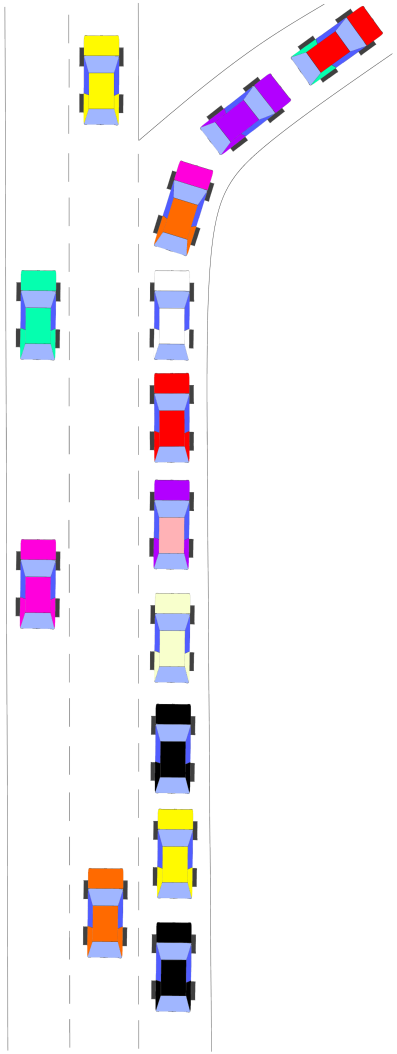
Научный руководитель: к.ф.-м.н. Чехович Юрий Викторович

Московский физико-технический институт

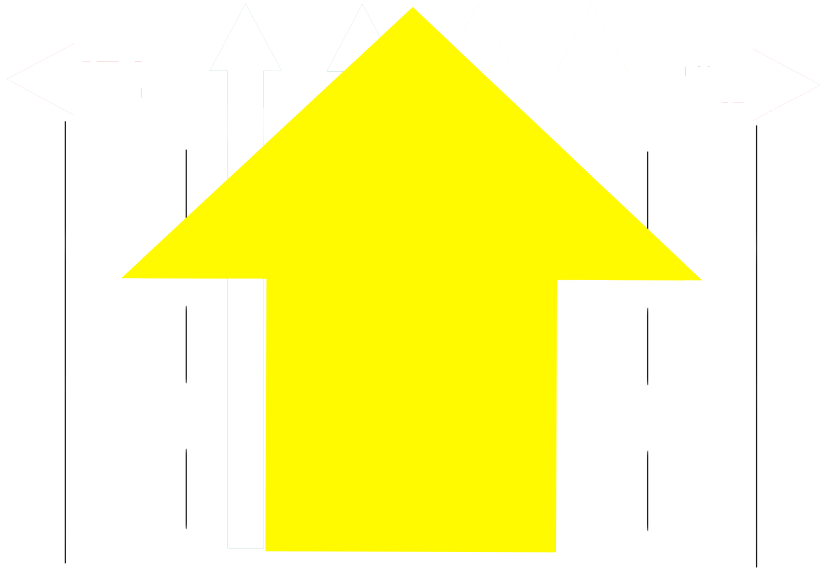
Сколковский институт науки и технологий

Введение

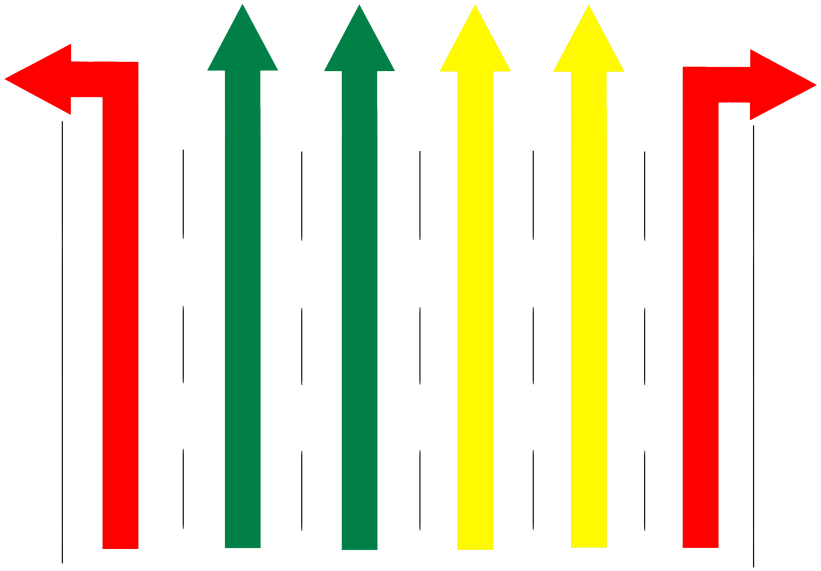




Введение



Текущий подход



Желаемый результат

Данные

Имеются GPS-треки и скорости автомобилей, проехавших по дороге, но неизвестна полоса их движения.

Введение

Цель

Учесть неоднородность транспортного потока по полосам движения для построения оптимальных маршрутов.

Проблема

Точности GPS-устройств не достаточно для отслеживания полосы движения автомобиля.

Решение

Оценка неоднородности потока методами анализа данных дорожного движения.

Литература

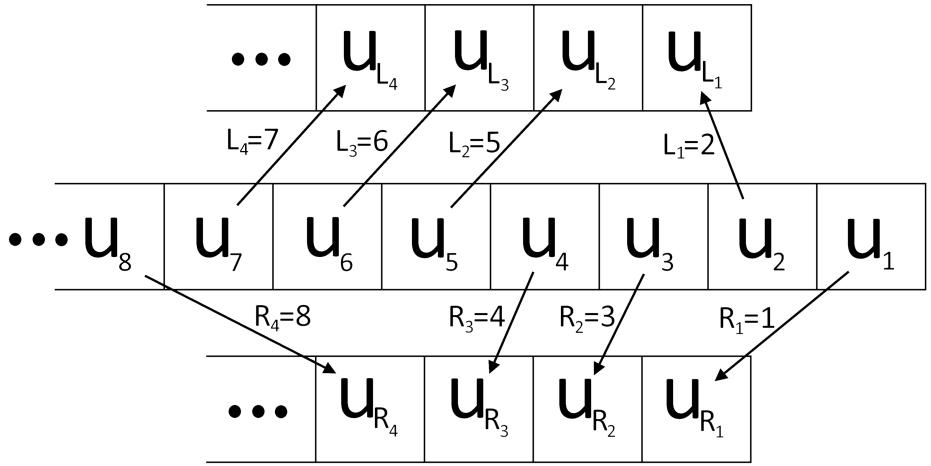
- Bo Xu; Tiffany Barkley; Andrew Lewis; Jane MacFarlane; Davide Pietrobon and Matei Stroila.
  Real-time detection and classification of traffic jams from probe data.
  Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 79, 2016, ACM.
- Robert W Thomas and José M Vidal.
  Toward detecting accidents with already available passive traffic information.
  Computing and Communication Workshop and Conference (CCWC), 2017 IEEE 7th Annual, 1–4, 2017, IEEE.
- Juan Susan Pan; Iulian Sandu Popa and Cristian Borcea.
  Divert: A distributed vehicular traffic re-routing system for congestion avoidance.
  IEEE Transactions on Mobile Computing, 16, 1, 58–72, 2017.

Предлагаемый подход

Разбить процесс решения задачи на две части:

1. **Статистическую:** оценивать вероятность проезда конкретного автомобиля по конкретной полосе движения.
2. **Оптимизационная:** искать наиболее вероятную среднюю скорость полосы движения исходя из решения статистической части.

Оптимизация: расщепление последовательности



Дано:

- $\{u_i\}_{i=1}^n$ — последовательность чисел
- $\{L_t\}_{t=1}^T$ — классы, на которые разбивается $\{u_i\}_{i=1}^n$
- P — матрица $n \times T$ распределения u_i по L_t
- $A(L_t)$ — статистика, заданная на множестве подпоследовательностей $\{u_i\}_{i=1}^n$.

Найти:

Оценки величин $\{A(L_t)\}_{t=1}^T$

Рассматриваемая статистика

В данной работе рассматривается расщепление на два потока и статистика экспоненциального скользящего среднего:

$$A(L) = \lambda \sum_{i=1}^{|L|} (1-\lambda)^{i-1} u_{L_i}.$$

Мера схожести статистик вводится в виде:

$$F = \sum_{D=L,R} \alpha_D (A(D) - A(K_D))^2, \text{ где } \alpha_D > 0.$$

Здесь K_D — истиная подпоследовательность индексов, а D — ее приближение.

Задача стохастического программирования

Задача:

$$\mathbb{E}\{F|D\} \longrightarrow \min_{D}$$

В результате серии преобразований, получаем эквивалентную постановку задачи:

$$x^* = \operatorname*{argmin}_{x \in \mathcal{X}} x^T \hat{u}\hat{u}^T x - 2(1-\alpha)\big(\mathbb{1} - D(p)^{-1}\pi(p,\lambda)\big)^T \hat{u}\hat{u}^T x.$$

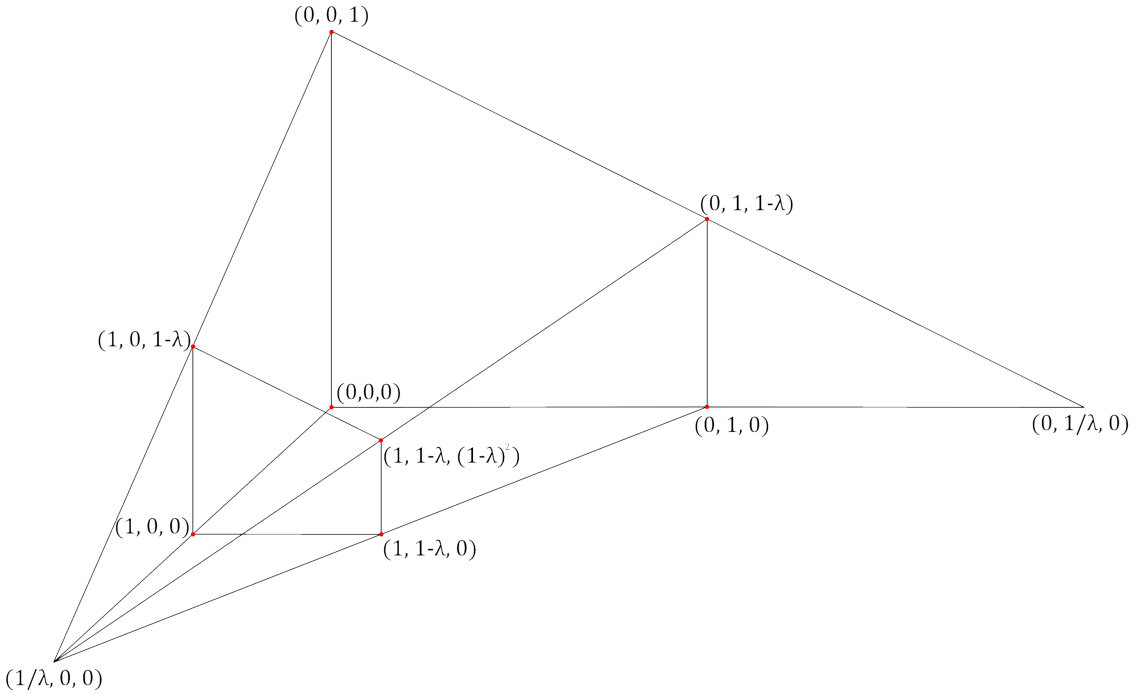
$$\mathcal{X} = \left\{ x \middle| \exists d, \exists [k_1, \ldots, k_d] : x^k = \begin{cases} 0 & k \notin [k_1, \ldots, k_d] \\ (1-\lambda)^{m-1} & k = k_m \end{cases} \right\}$$

Итоговый вид задачи

$$|x^T \hat{u} - a| \longrightarrow \min_{x \in \mathcal{X}}$$

$$\mathcal{X} = \text{extr}\left(\{x \in \mathbb{R}^{n+2} \big| \lambda N^T x \leq \mathbb{1}; x \geq 0\}\right),$$

где $a = a(\hat{u}, \lambda, P, \alpha), \quad N = N(\lambda).$



Замечания

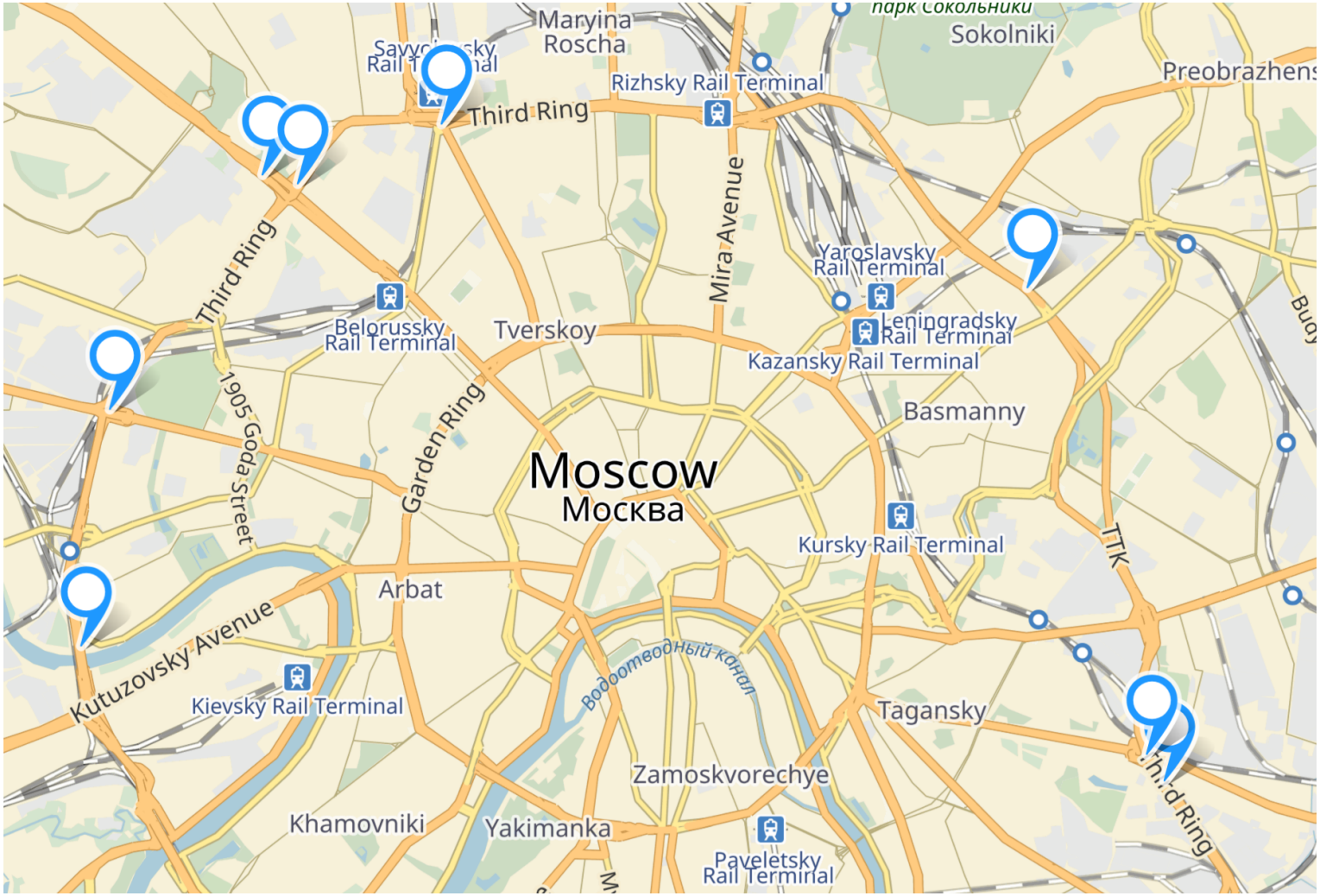
- Данная задача сводится к задаче линейной смешанной целочисленной оптимизации.
- Функцию a сложно посчитать даже при известных параметрах задачи.
- К оценке матрицы P в контексте транспортных систем возможны различные подходы.
- P, состоящая из нулей и единиц избавляет нас от решения задачи оптимизации.

Применение техник классификации

Приближение двухстолбцовой матрицы P из нулей и единиц при помощи классификации:

$$\|P - \hat{P}\|_F^2 \propto \frac{\sum_{i=1}^{n} \mathbb{I}(\text{estimated_class}_i = \text{real_class}_i)}{n} = Accuracy$$

Применение техник классификации: *i.i.d.*



Применение техник классификации: *i.i.d.*

Таблица: Результаты работы *i.i.d.* алгоритмов классификации

	SVM	**RandFor**	**BagKNN**	**GradBst**	**LogReg**	**NB**	**NNet**
1	0.722	0.879	0.873	0.886	0.717	0.727	0.880
2	0.750	0.857	0.866	0.865	0.768	0.860	0.864
3	0.744	0.751	0.765	0.765	0.731	0.548	0.796
4	0.907	0.920	0.929	0.925	0.912	0.715	0.925
5	0.769	0.799	0.799	0.801	0.761	0.722	0.800
6	0.711	0.772	0.777	0.778	0.703	0.692	0.769
7	NA	0.764	0.780	NA	NA	0.697	0.789
8	0.725	0.742	0.749	0.753	0.750	0.690	0.751

Применение техник классификации: онлайн

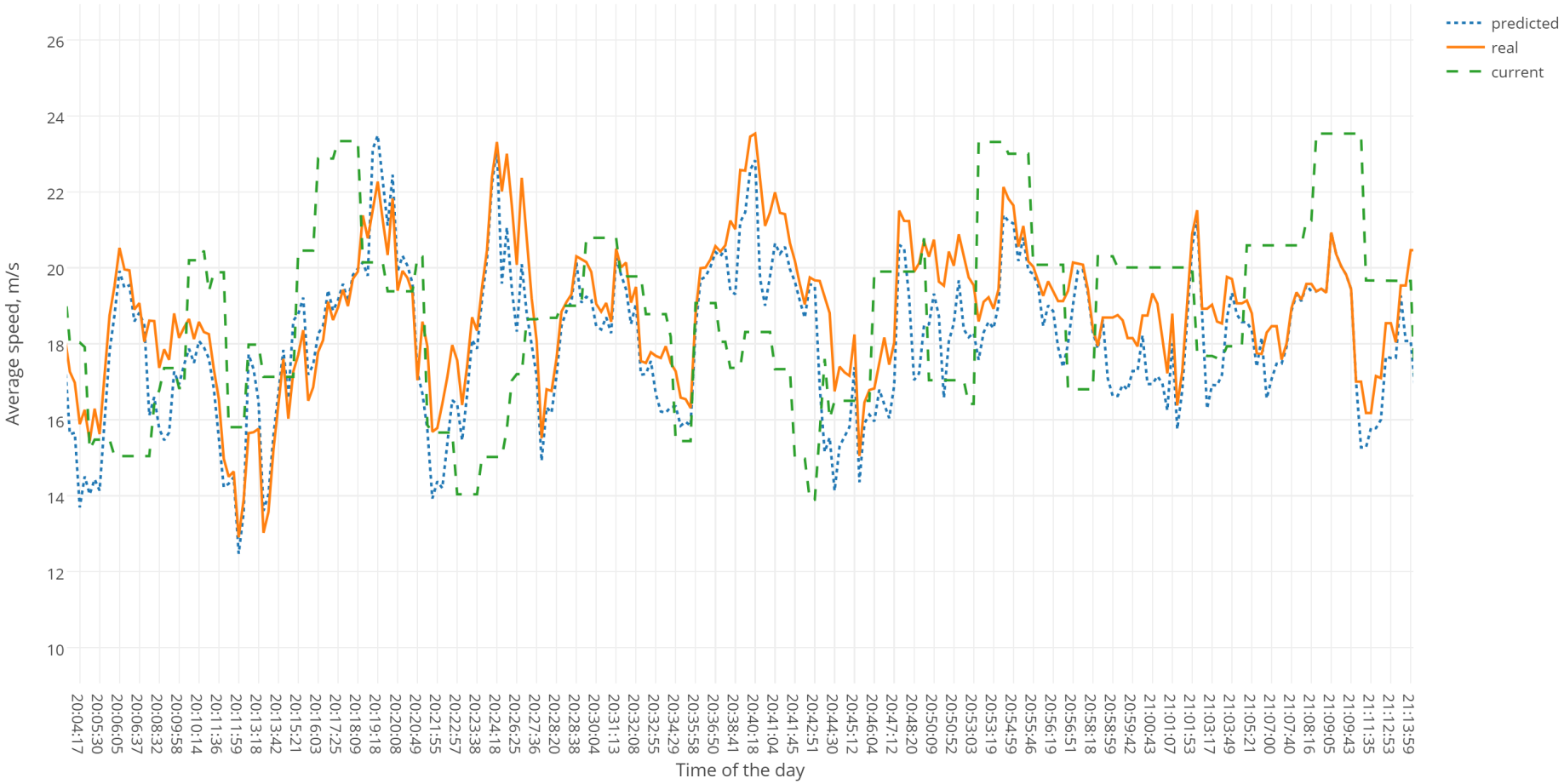


Иллюстрация работы онлайн-алгоритма

Заключение

- Разработана формальная постановка прикладной задачи.
- Задача разделена на две части: оптимизационную и статистическую.
- Постановка оптимизационной части сведена к виду, удобному для релаксации и применений стандартных алгоритмов оптимизации.
- Проанализированы возможные пути решения статистической части задачи.
- Один из этих путей реализован на практике, показаны его приимущества и недостатки.