АНАЛИЗ БОЛЬШИХ ДАННЫХ В ГЕОИНФОРМАЦИОННОЙ ЗАДАЧЕ КРАТКОСРОЧНОГО ПРОГНОЗИРОВАНИЯ ПАРАМЕТРОВ ТРАНСПОРТНОГО ПОТОКА НА БАЗЕ МЕТОДА *K* БЛИЖАЙШИХ СОСЕДЕЙ

А.А. Агафонов [1], А.С. Юмаганов [1], В.В. Мясников [1,2]

[1] Самарский национальный исследовательский университет имени академика С.П. Королёва, 443086, Россия, г. Самара, Московское шоссе д. 34,

[2]ИСОИ РАН – филиал ФНИЦ «Кристаллография и фотоника» РАН, 443001, Россия, г. Самара, ул. Молодогвардейская, д. 151

Аннотация

Точная и своевременная информация о текущем и прогнозном распределении транспортных потоков является важным фактором функционирования интеллектуальных транспортных систем. Использование этих данных позволит транспортным агентствам эффективнее решать задачу управления трафиком, участникам дорожного движения точнее планировать маршрут поездки и снизить время движения, и в целом повысит эффективность использования транспортной инфраструктуры. В данной статье представлена модель краткосрочного прогнозирования трафика, основанная на методе *k* ближайших соседей, которая учитывает пространственное и временное распределение транспортных потоков. Разработанная модель реализована с помощью фреймворка Apache Spark на основе модели распределённых вычислений MapReduce. Экспериментальные исследования представленной модели по данным о распределении транспортных потоков в транспортной сети города Самары позволяет сделать вывод, что предлагаемая модель обладает высокой точностью прогнозирования и временем работы, достаточным для прогнозирования в режиме реального времени.

Ключевые слова: транспортный поток, краткосрочное прогнозирование, *k* ближайших соседей, MapReduce.

Цитирование: **Агафонов, А.А.** Анализ больших данных в геоинформационной задаче краткосрочного прогнозирования параметров транспортного потока на базе метода *k* ближайших соседей / А.А. Агафонов, А.С. Юмаганов, В.В. Мясников // Компьютерная оптика. – 2018. – Т. 42, № 6. – С. 1101-1111. – DOI: 10.18287/2412-6179-2018-42-6-1101-1111.

Введение

Проблемы организации дорожного движения являются общими для всех крупных городов. Дорожные заторы приводят к социальным, экономическим и экологическим проблемам, что обуславливает первостепенную значимость задач транспортного планирования и логистики. Для решения этих задач важную роль играет получение точной и своевременной информации о текущем и прогнозном распределении транспортных потоков, поэтому задача прогнозирования дорожного движения является предметом активных научных и практических исследований.

Обычно выделяют несколько направлений решения проблемы дорожных заторов: модификация транспортной инфраструктуры, развитие пассажирского транспорта и управление транспортными потоками. Развитие первых двух направлений часто ограничено бюджетными или социальными факторами, в то время как решение задачи управления транспортными потоками постоянно совершенствуется благодаря развитию технологий сбора и обработки данных о параметрах транспортных потоков (скорость, плотность).

В последнее время большое внимание исследователей уделяется методам и алгоритмам, управляемым данными. Такой интерес обусловлен разработкой новых технологий, методов и программного обеспечения для обработки огромных массивов данных в рамках парадигмы «Big Data», наличием нескольких источников данных для прогнозирования транспортных потоков, а также распространением концепции «открытых данных», подразумевающей свободное распространение определённых данных для машиночитаемого использования.

Краткосрочное прогнозирование трафика решает задачу прогнозирования состояния транспортных потоков на основе текущей и архивной информации о параметрах транспортных потоков. Обзор последних достижений в области прогнозирования трафика, а также основных нерешённых технических проблем можно найти в работе [1]. Более подробный обзор существующих методов и алгоритмов прогнозирования представлен в следующем параграфе.

Данная работа опирается на один из главных непараметрических методов краткосрочного прогнозирования движения – метод *k* ближайших соседей (kNN). Результаты, представленные в [2, 3, 4], показали, что kNN превосходит другие современные сопоставимые модели, включая ANN, SARIMA, random forest и Naïve Bayes. Однако, если объём выборки данных достаточно велик, kNN может быть неподходящим для прогнозирования в режиме реального времени из-за больших вычислительных затрат. Несмотря на это, краткосрочному прогнозированию транспортных потоков с точки зрения обработки больших данных с использованием модели распределённых вычислений, в частности, с использованием модели MapReduce, посвящено относительно небольшое число работ [4, 5].

В данной статье мы решаем задачу краткосрочного прогнозирования на горизонт прогноза в 10 минут. Основной упор делается на создание распределённой модели прогнозирования на основе взвешенного алгоритма kNN с учётом пространственных и времен-

ных характеристик транспортных потоков в пространственно компактной области транспортной сети. Для распределённой обработки данных используется двухуровневая модель MapReduce, реализованная в составе фреймворка с открытым исходным кодом Apache Spark. Экспериментальный анализ данных о движении транспортных средств позволяет сделать вывод, что предлагаемая модель обладает высокой точностью прогнозирования и временем выполнения, достаточным для прогнозирования в режиме реального времени.

Основные этапы решения задачи прогнозирования и вклад этой работы сводится к следующему:

- Для решения задачи прогнозирования используется распределённый взвешенный метод kNN. Учитываются пространственные и временные характеристики транспортных потоков. В отличие от работы [4], для построения используются не только смежные сегменты дорожной сети, а характеристики транспортных потоков в компактном пространственном кластере.
- Для увеличения быстродействия обработки данных выполняется снижение размерности вектора признаков с пространственно-временным описанием кластера транспортной сети путём устранения зависимости между значениями параметров потоков с помощью метода главных компонент.
- Для снижения вычислительных затрат предлагается алгоритм распределённой обработки большого объёма данных с помощью фреймворка Apache Spark, реализующего модель распределённых вычислений MapReduce.

Статья организована следующим образом. В первом параграфе кратко приводится обзор литературы по смежным с настоящей работой темам. Во втором параграфе приводится формулировка проблемы. Предлагаемая модель и её распределённая реализация описаны в параграфах 3 и 5. В четвёртом параграфе описаны используемые в работе способы разбиения графа транспортной сети на подграфы (кластеры). В шестом параграфе представлены постановка и результаты экспериментальных исследований предложенной модели для проверки точности, эффективности и масштабируемости предложенного подхода. В завершение работы представлены заключение и возможные направления дальнейших исследований.

1. Обзор литературы

Краткосрочное прогнозирование трафика

Краткосрочное прогнозирование трафика решает задачу прогнозирования состояния транспортных потоков на основе текущей и архивной информации о параметрах транспортных потоков. Большинство исследований по этой тематике сосредоточены на разработке методов для моделирования характеристик транспортных потоков (например, плотности или скорости потока). Обзор методов краткосрочного прогнозирования трафика представлен в [6]. В статье [7] проведён анализ современного состояния исследований краткосрочного прогнозирования параметров транспортных потоков по различным критериям, включая используемые технологии сбора данных, модели прогнозирования, прогнозируемые характеристики транспортных потоков и т.д.

Учитывая неоднородность и динамические свойства транспортных потоков, а также сложные нелинейные взаимодействия участников дорожного движения друг с другом и с транспортной инфраструктурой, можно говорить о высокой степени изменчивости характеристик транспортных потоков в сети. Кроме того, состояние транспортных потоков на определённом сегменте сети сильно зависит от состояния потоков на смежных с ним сегментах, что затрудняет прогнозирование, особенно для краткосрочного интервала времени. Очевидно, что краткосрочное прогнозирование параметров транспортных потоков является сложной задачей, и, как следствие, для её решения в течение последних десятилетий было предложено много моделей, методов и алгоритмов. Известные подходы к прогнозированию используют различные эмпирические и теоретически методы, которые в целом могут быть классифицированы по трём категориям:

1. Параметрические методы [6, 8], включая модели временных рядов (модель авторегрессии скользящего среднего ARMA, интегрированную модель ARIMA, а также её модификацию с учётом сезонности SARIMA [9], векторную модель авторегрессии VARMA [10]), модели пространства состояний (например, фильтр Калмана [11, 12]), модели на основе динамического распределения трафика (DTA) [13] и т.д.

2. Непараметрические методы [2], включая модели искусственных нейронных сетей [14, 15], метод k ближайших соседей (kNN) [16, 17], метод опорных векторов [18, 19], модель байесовской сети [20, 21].

3. Гибридные методы, сочетающие параметрические и непараметрические методы [22, 23], например, комбинация модели ARIMA с другими моделями для повышения точности прогноза [24], объединение статистических методов и моделей нейронных сетей [25], а также комбинация моделей прогнозирования с методами предобработки данных [26, 27].

Все описанные методы имеют как преимущества, так и недостатки при работе в разных условиях, поэтому сложно сделать вывод, что один метод значительно превосходит другие во всех режимах прогнозирования. Это объясняется в том числе и тем фактом, что точность моделей прогнозирования, которые обучаются и проверяются на небольших специальных наборах данных, зависит и от характеристик транспортных потоков в используемых выборках [5]. Кроме того, большинство существующих методов работают в автономном режиме, и поэтому имеют ограничения на используемые вычислительные ресурсы, а также возможности хранения и обработки данных.

В данной статье разрабатывается архитектура распределённой системы краткосрочного прогнозирования параметров транспортных потоков с исполь-

зованием модели MapReduce, реализованной в составе фреймворка Apache Spark, для эффективной обработки данных больших объёмов.

Методы пространственного разбиения графа транспортной сети

Задаче кластеризации данных посвящено большое число работ, обзор основных методов и алгоритмов кластеризации приведён в [28]. Алгоритмы кластеризации используются в различных областях, включая анализ данных [29], сегментацию изображений [30] или информационный поиск [31].

Однако кластеризация транспортных сетей обладает своими особенностями из-за динамических свойств характеристик транспортных потоков. В частности, алгоритмы кластеризации транспортных сетей должны удовлетворять следующим критериям:

1) малая дисперсия значений плотности транспортных потоков внутри каждого кластера;
2) пространственно-близкие компактные формы кластеров.

В некоторых задачах управления требуется также удовлетворение требования малого количества кластеров, что позволит разрабатывать стратегии управления без необходимости детализированного описания потоков между различными кластерами.

В работе [32] изучается задача кластеризации транспортных сетей на основе информации о характеристиках транспортных потоков в течение определённого периода времени для оценки фундаментальной диаграммы транспортных потоков. Для разбиения транспортной сети на несколько однородных регионов в статье [32] используется алгоритм Normalized Cut (NCut), который эффективно выделяет главные компоненты и гарантирует пространственную компактность форм сегментов. Схожие принципы для кластеризации применяются в алгоритме Min-Max Cut, который отличается видом используемой целевой функцией, но также стремится минимизировать дисперсию внутри кластера и максимизировать дисперсию между кластерами [33].

2. Формулировка проблемы

Улично-дорожную сеть будем рассматривать как ориентированный граф $G=(V,E)$, в котором вершины V, $N_V=|\mathrm{V}|$ соответствуют перекрёсткам дорожной сети, рёбра E, $N_E=|\mathrm{E}|$ соответствуют сегментам дорожной сети между перекрёстками.

Обозначим V_t^j – параметр транспортного потока на сегменте $j\in E$ в момент времени t. В качестве параметра транспортного потока могут выступать следующие величины:

- средняя скорость транспортного потока,
- плотность потока,
- поток (собственно величина потока).

В настоящей работе в качестве прогнозируемого параметра транспортного потока в экспериментальных исследованиях используется средняя скорость движения.

Учитывая введённые обозначения, формальная постановка задачи получения краткосрочного прогноза для заданного параметра транспортного потока может быть сделана следующим образом:

Имея заданный граф $G=(V,E)$ *и последовательность* V_t^j, $j\in E$, $t=1,2,\ldots,T$ *наблюдаемых значений параметров транспортных потоков, рассчитать оценку (спрогнозировать) параметров в момент времени* $(t+\Delta)$ *для определённого горизонта прогноза* Δ.

3. Предложенная модель

Предлагаемое в работе решение задачи краткосрочного прогнозирования транспортных потоков основывается на непараметрическом регрессионном методе *k* ближайших соседей.

Для применения метода необходимо решить следующие задачи:

1) определить вектор признакового описания транспортного потока;
2) определить подходящую метрику расстояния для определения близости между векторами признаков, описывающих текущие характеристики транспортных потоков и архивные данные;
3) определить функцию вычисления прогноза по набору ближайших соседей.

Решение каждой задачи описано в следующих подпараграфах.

Вектор признаков

Выбор вида вектора признаков в методе *k* ближайших соседей зависит от конкретного приложения метода. Для решения задачи прогнозирования транспортных потоков целесообразно использовать вектор признаков, учитывающий пространственные и временные корреляции параметров транспортных потоков.

В работе [4] в качестве вектора признаков предложено использовать значение параметров транспортных потоков на текущем, предыдущем и следующем сегменте дорожной сети за T временных интервалов:

$$(V_{t-T}^j,\ldots,V_{t-1}^j,V_t^j,V_{t-T}^{j-1},\ldots,V_{t-1}^{j-1},V_t^{j-1},\\ V_{t-T}^{j+1},\ldots,V_{t-1}^{j+1},V_t^{j+1}). \quad (1)$$

Однако такое описание не учитывает транспортную ситуацию на смежных сегментах. Кроме того, в некоторых случаях предыдущий / следующий дорожный сегмент не может быть определён однозначно. Поэтому для описания транспортных потоков предлагается формировать вектор признаков с учётом характеристик транспортных потоков в пространственно-компактном кластере графа дорожной сети.

В настоящей работе используется следующий способ формирования вектора признаков:

1. Граф улично-дорожной сети разбивается на пространственно компактные непересекающиеся кластеры $\{G_i\}$. В каждом кластере формируется вектор признаков:

$$\{V_t^j\},\ j\in G_i,\ t=t_{cur}-T,\ldots,t_{cur}. \quad (2)$$

2. Выполняется снижение размерности исходного вектора признаков путём устранения пространственно-временной зависимости значений параметров потоков $\{X_n\}^i$, $n=1,\ldots,N$.
3. Результирующий вектор признаков для каждого сегмента $j\in E$ формируется из исходного вектора признаков для сегмента j и вектора признаков для кластера графа i, которому принадлежит сегмент:

$$S_j = (\{V_t^j\},\{X_n\}^i),\ j\in G_i,\quad t=t_{cur}-T,\ldots,t_{cur},\ n=1,\ldots,N. \tag{3}$$

Способы разбиения графа транспортной сети на подграф и выделения кластеров дорожной сети подробно описаны в параграфе 4.

Мера близости

Для определения близости между векторами признаков необходимо определить подходящую метрику расстояния. В литературе описаны разные варианты определения меры расстояния между векторами, в т.ч. евклидово расстояние, расстояние Махаланобиса, расстояние Хэмминга.

В данной работе используется взвешенное евклидово расстояние с учётом тренда, предложенное в работе [4], модифицированное для использования вектора признаков, описывающего кластеры транспортной сети $\{X\}$:

$$d(S,\overline{S}^i) = d^{link}(V,\overline{V}^i) + \gamma d^{pca}(X,\overline{X}^i), \tag{4}$$

$$d^{link}(V,\overline{V}^i) = \alpha\sqrt{\sum_{t=1}^{T}\beta^{T-t+1}(V_t-\overline{V}_t^i)^2} + (1-\alpha)\sqrt{\sum_{t=2}^{T}\sum_{\delta=1}^{t-1}((V_t-V_\delta)-(\overline{V}_t^i-\overline{V}_\delta^i))^2}, \tag{5}$$

$$d^{pca}(X,\overline{X}^i) = \sqrt{\sum_{n=1}^{N}(X_n-\overline{X}_n^i)^2}, \tag{6}$$

где $0\le\alpha\le 1$, $0<\beta\le 1$, $0\le\gamma\le 1$ – коэффициенты, T – количество временных интервалов в векторе признаков, N – количество элементов в векторе признаков, описывающем кластер, $d(V,\overline{V}^i)$ – расстояние между вектором признаков, описывающим текущее распределение транспортных потоков, и i-м архивным вектором признаков, S – значение вектора признаков, описывающего текущий транспортный поток, $\overline{S}_i$ – значение вектора признаков, описывающего i-й архивный транспортный поток, $V_t, \overline{V}_t^i$ – значения векторов признаков, описывающих соответственно текущие и архивные значения транспортного потока на заданном дорожном сегменте за временной интервал t, $X_n, \overline{X}_n{}^i$ – n-е значения векторов признаков, описывающих соответственно текущее и архивное состояние транспортного потока в кластере.

Функция прогнозирования

Традиционным подходом для оценки значения при использовании метода k ближайших соседей в задаче регрессии является выбор среднего или средневзвешенного значения по k ближайшим к оцениваемому вектору признаков [2].

Функция прогнозирования по средневзвешенному значению ближайших векторов признаков имеет вид:

$$\hat{X}_{T+1} = \sum_{k=1}^{K}\frac{d_k^{-1}}{\sum_{k=1}^{K}d_k^{-1}}X_{T+1}^k, \tag{7}$$

где $\hat{X}_{T+1}$ – прогнозное значение параметра транспортного потока в момент времени $T+1$, X_{T+1}^k – значение параметра транспортного потока k-го ближайшего соседа в соответствующий момент времени, d_k – расстояние между вектором признаков, описывающим текущее состояние транспортных потоков, и k-м ближайшим соседом.

В работе используется комбинированная функция прогнозирования, учитывающая средневзвешенное значение векторов признаков и тренд прогноза:

$$\hat{X}_{T+1} = \theta\sum_{k=1}^{K}\frac{d_k^{-1}}{\sum_{k=1}^{K}d_k^{-1}} + (1-\theta)\left(X_T + \frac{1}{KT}\sum_{k=1}^{K}\sum_{t=1}^{T}(X^k{}_{T+1} - X_t^k)\right), \tag{8}$$

где $0\le\theta\le 1$ – коэффициент, K – количество ближайших соседей.

4. Разбиение графа на подграфы

В работе используются несколько способов кластеризации транспортной сети:

- кластеризация по территориальному признаку: в один кластер попадают рёбра графа, соответствующие сегментам дорожной сети, принадлежащим указанной прямоугольной области;
- кластеризация с учётом матрицы расстояний: для каждого ребра графа в соответствующий ему кластер попадают рёбра, находящиеся на расстоянии от выбранного ребра, не превышающем заданное;
- кластеризация на основе информации о характеристиках транспортных потоков в течение определённого периода времени.

Кластеризация по территориальному признаку

Введём дополнительные обозначения. Пусть ребру графа $i\in E$ соответствует сегмент дорожной сети e_i с известными координатами начала и конца сегмента $x_{start}=\left(x_{start}^0, x_{start}^1\right)$ и $x_{end}=\left(x_{end}^0, x_{end}^1\right)$ соответственно.

Тогда кластеризация графа по территориальному признаку может быть описана следующим образом:

1. Выбирается число формируемых подграфов в количестве M_0, M_1.

2. В подграф G_m^{bbox} с номером $m=m_0M_1+m_1$, $\left(m_0=\overline{0,M_0-1}; m_1=\overline{0,M_1-1}\right)$ относятся те рёбра $i\in E$, координаты одной из вершин соответствующих дорожных сегментов которых по-

падают в соответствующую прямоугольную область Π_{m_0,m_1}

$$G_m^{bbox} \equiv \left\{ i \in E : \quad x_{start}^i \in \Pi_{m_0,m_1} \vee x_{end}^i \in \Pi_{m_0,m_1} \right\},$$

где

$$\Pi_{m_0,m_1} \equiv \left[x_{\min}^o + \frac{m_0}{M_0}\Delta^0, x_{\min}^0 + \frac{m_0+1}{M_0}\Delta^0 \right] \times \times \left[x_{\min}^1 + \frac{m_0}{M_0}\Delta^1, x_{\min}^1 + \frac{m_1+1}{M_1}\Delta^1 \right],$$

$$x_{\min}^s = \min_{\substack{\upsilon=\{start,end\} \\ i \in E}} x_\upsilon^{s,i}, \quad x_{\max}^s = \max_{\substack{\upsilon=\{start,end\} \\ i \in E}} x_\upsilon^{s,i},$$

$$\Delta^s = x_{\max}^s - x_{\min}^s, \qquad s = 0,1.$$

Число подграфов по вертикали и горизонтали M_0, M_1 выбирается эмпирически. Будем считать, что каждый сегмент дорожной сети может попасть только в один кластер.

Данный способ обладает следующими недостатками:

1) прогноз характеристик транспортных потоков на граничных сегментах дорожной сети может быть неточным;
2) при формировании кластеров не учитываются характеристики транспортных потоков.

В следующих подпараграфах рассмотрены два способа выделения подграфов, исправляющих указанные недостатки.

Кластеризация с учётом матрицы расстояний

Определим расстояние $r(i,j)$ между двумя рёбрами графа $i \in E$ и $j \in E$ как длину кратчайшего пути от начальной вершины ребра i, x_{start}^i до конечной вершины ребра j, x_{end}^j в невзвешенном графе (т.е. количество рёбер в кратчайшем пути). Расстояние между всеми вершинами графа определяется на основе матрицы смежности графа.

Тогда каждому ребру графа $i \in E$ ставится в соответствие кластер G_i^{dist} по следующему правилу:

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

где R – эмпирически выбираемое максимальное расстояние, определяющее размер кластера.

При данном разбиении разные кластеры транспортной сети будут содержать общие рёбра графа.

Кластеризация с учётом транспортных потоков

Для кластеризации транспортной сети на несколько однородных регионов с учётом транспортных потоков используется алгоритм Normalized Cut (NCut), который позволяет эффективно выделять главные компоненты и гарантирует пространственную компактность форм сегментов.

Нахождение точного минимума целевой функции в алгоритме NCut является NP-полной задачей, однако дискретное решение может быть аппроксимировано вещественным путём решения обобщённой задачи поиска собственных значений. Кластеризация транспортной сети с помощью алгоритма NCut может быть описана следующим образом:

1. Используя граф транспортной сети, установить вес ребра $w(i,j)$ как меру сходства двух сегментов дорожной сети $i \in E, j \in E$:
$$w(i,j) = \begin{cases} \exp(-(V_i - V_j)^2), & r(i,j) < R_{\max}; \\ 0 \qquad\qquad\quad , & \textit{иначе.} \end{cases}$$
2. Решить эквивалентную систему собственных векторов для второго наименьшего собственного значения.
3. Дискретизировать собственный вектор, соответствующий второму наименьшему собственному значению, разделить граф на два подграфа.
4. Повторить процесс рекурсивно для отдельных подграфов, если необходимо.

Результат кластеризации обозначим как $\{G_i^{flow}\}$.

В следующем параграфе описана реализация представленной в данной работе модели краткосрочного прогнозирования транспортного потока с использованием модели распределённых вычислений MapReduce.

5. Реализация в MapReduce

В процессе решения задачи прогнозирования транспортного потока с помощью представленной в данной работе модели используется большой объём архивных и текущих данных. Для повышения эффективности производимых вычислений, использующих такой объём информации, предлагаемая модель была реализована с помощью фреймворка Apache Spark [34], использующего модель распределённых вычислений MapReduce [35]

Apache Spark – фреймворк с открытым исходным кодом для реализации распределённой обработки неструктурированных и слабоструктурированных данных, входящий в экосистему проектов Hadoop. В отличие от классического обработчика из ядра Hadoop, реализующего двухуровневую модель, MapReduce с дисковым хранилищем использует специализированные примитивы для рекуррентной обработки в оперативной памяти, благодаря чему позволяет получать значительный выигрыш в скорости работы для некоторых классов задач, в частности, возможность многократного доступа к загруженным в память пользовательским данным делает библиотеку привлекательной для алгоритмов машинного обучения.

В рамках модели MapReduce обработка данных происходит параллельно на нескольких вычислительных узлах. Работа MapReduce состоит из трёх основных этапов: Map, Shuffle и Reduce. На рис. 1 представлена схема работы предлагаемой модели на основе MapReduce.

Как видно из рис. 1, на первом шаге осуществляется подготовка входных данных для Map-этапа. Сначала происходит разбиение исторических и тестовых данных на разделы. Оптимальное количество таких разделов зависит от объёма обрабатываемой информации и количества вычислительных узлов. Затем формируются упорядоченные пары разделов исторических и тестовых данных с помо-

щью декартова произведения. Далее на Map-шаге к каждой паре разделов применяется map-функция, которая возвращает промежуточный набор пар ключ/значение – тестовый элемент/локальный список k ближайших соседей. На шаге Shuffle осуществляется группировка пар ключ-значение и их передача функции reduce. На заключительном Reduce-шаге для каждого элемента тестовых данных множество списков локальных k ближайших соседей преобразуется в результирующий (глобальный) список k ближайших соседей. Полученные списки k ближайших соседей впоследствии используются для нахождения прогнозируемой величины транспортного потока.

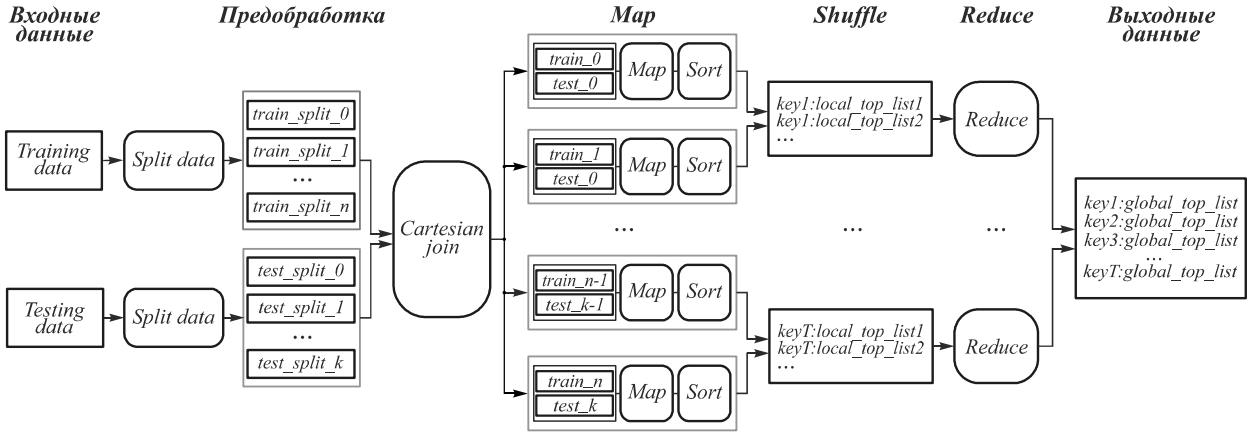

Рис. 1. Схема работы предлагаемой модели на основе MapReduce

Результаты оценки эффективности и масштабируемости предлагаемой модели прогнозирования на основе модели распределённых вычислений MapReduce представлены в параграфе 6.

6. Экспериментальные исследования

Экспериментальные исследования разработанной модели проводились для транспортной сети г. Самары, граф которой включает в себя 26018 дорожных сегментов. В качестве исходных данных для проведения экспериментальных исследований использовались значения средней скорости движения транспортного потока (в км/ч), полученные за 50 дней, начиная с 27 июля 2018 г.

В ходе проведения экспериментальных исследований было произведено сравнение представленной модели при использовании различных алгоритмов разбиения графа, сезонной модели временных рядов SARIMA и модели TDUD-KNN, представленной в работе [4]. В основе модели TDUD-KNN также лежит метод k ближайших соседей с вектором признаков (1), который содержит в себе информацию о транспортном потоке на заданном и соседних сегментах. Исследования качества прогноза данных моделей проводились для каждого дня исходных данных, при этом в качестве архивных данных выступали данные о средней скорости транспортного потока, полученные за весь названный выше период времени (50 дней), за исключением рассматриваемого дня.

Для оценки качества прогноза представленной модели использовались следующие метрики: средняя абсолютная ошибка (MAE) и средняя абсолютная ошибка в процентах (MAPE):

$$MAE = \frac{1}{n}\sum_{t=1}^{n} | V_t - \hat{V}_t | , \qquad (9)$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} \frac{| V_t - \hat{V}_t |}{V_t} \times 100\% , \qquad (10)$$

где V_t – реальное значение величины транспортного потока на временном интервале t, $\hat{V}_t$ – спрогнозированное значение величины транспортного потока на временном интервале t, n – общее количество прогнозов.

В первую очередь были проведены эксперименты по выбору параметра k (числа соседей) в предложенной модели. Эксперименты проводились на части выборки за 30 дней. Зависимость ошибки MAPE от параметра k представлена на рис. 2.

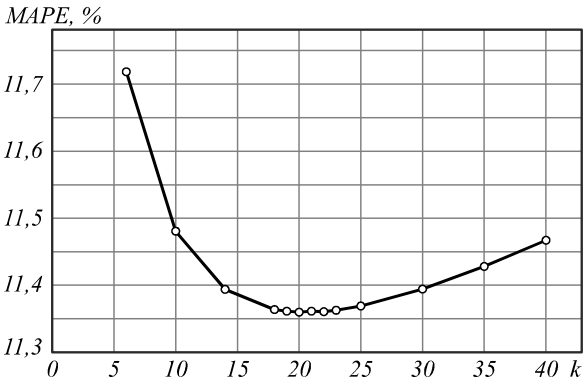

Рис. 2. Зависимость MAPE от числа соседей k

Лучший результат был показан для значения $k=20$, далее в экспериментах использовалось это значение.

В табл. 1 представлено сравнение ошибок MAE и MAPE для представленной модели при различных способах кластеризации транспортной сети (G^{box}, G^{dist}, G^{flow}), модели SARIMA и модели TDUD-KNN.

Анализ представленных результатов показывает, что предлагаемая в данной работе модель даёт более качественный результат прогноза, чем известная ра-

нее модель. Согласно полученным результатам, все алгоритмы кластеризации графа показали схожее качество прогноза, с небольшим преимуществом лучший результат был показан при использовании алгоритма кластеризации графа транспортной сети с учётом информации о транспортных потоках.

Табл. 1. Сравнение моделей

Модель прогнозирования	MAE	MAPE
G^{box}	2,654	11,45
G^{dist}	2,653	11,42
G^{flow}	2,646	11,4
TDUD-KNN	2,732	11,76
SARIMA	2,677	11,64

На рис. 3, 4 представлены результаты оценки качества прогноза представленной модели с алгоритмом кластеризации с учётом транспортных потоков (NCut), моделей SARIMA и TDUD-KNN за две недели.

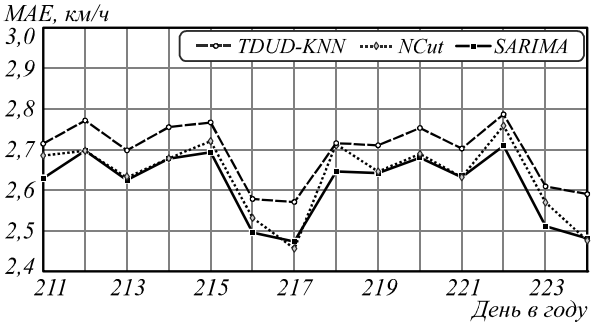

Рис. 3. Сравнение значений MAE представленной модели, моделей SARIMA и TDUD-KNN

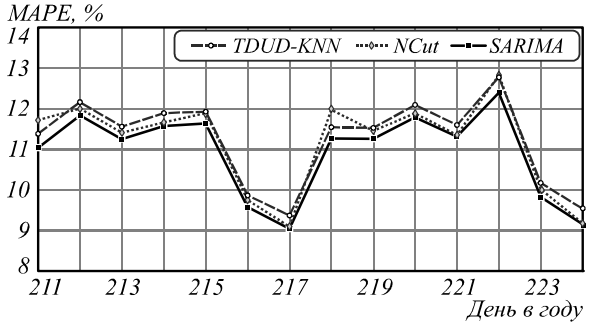

Рис. 4. Сравнение значений MAPE представленной модели, моделей SARIMA и TDUD-KNN

Как видно из рисунков, качество прогноза представленной модели превосходит качество прогноза ранее известной модели в каждый из рассматриваемых дней. Результаты представленной модели и модели SARIMA близки, однако точность предложенной модели может быть увеличена за счёт накопления большего объёма обучающей выборки.

Анализируя полученные результаты, можно сделать вывод, что в выходные дни качество прогноза заметно снижается при использовании каждой из рассматриваемых моделей. Это объясняется тем, что транспортный поток в выходные и рабочие дни существенно различается, в выходные дни транспортная ситуация стабильнее и предсказывается с большей точностью.

Для оценки масштабируемости и эффективности разработанной модели, в основе которой лежит модель распределённых вычислений MapReduce, был проведён ряд экспериментов. Вычисления проводились на кластере из шести компьютеров с CPU Core i5-3450, 8 Гб RAM. Для прогнозирования использовалась предложенная модель с учётом информации о транспортных потоках для кластеризации графа. Время работы модели усреднялось за 5 измерений.

Время работы предложенной модели (в секундах) для прогнозирования скорости транспортных потоков на всех сегментах дорожной сети для одного интервала времени в зависимости от числа вычислительных узлов показано в табл. 2.

Табл. 2. Сравнение времени работы модели

Число узлов	1	2	3	4	5	6
Время, с	346	176	139	101	88	74

Учитывая, что период обновления данных о состоянии транспортных потоков составляет 10 минут, можно сделать вывод, что модель позволяет прогнозировать состояние транспортных потоков в режиме реального времени.

Далее исследовалась масштабируемость предложенной модели, которая показывает, как меняется производительность системы при пропорциональном увеличении объёма обрабатываемой информации и мощности системы.

$$Scaleup = \frac{T_s}{T_{pp}}, \tag{11}$$

где T_s – время, затрачиваемое на решение заданной задачи на одном вычислительном узле; T_{pp} – время, затрачиваемое на решение задачи, размер входных данных которой увеличен в p раз, на p вычислительных узлах.

Чем ближе отношение (11) к единице, тем лучшей масштабируемостью обладает система. Результаты представлены на рис. 5.

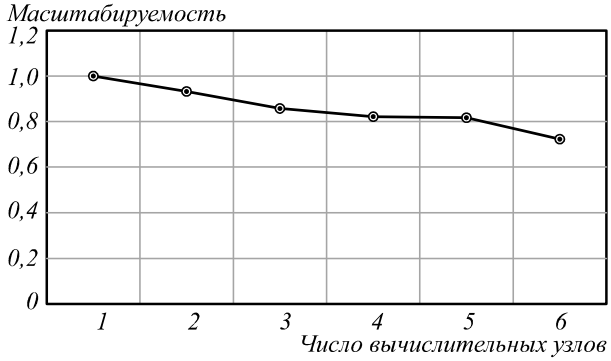

Рис. 5. Масштабируемость

На основе анализа полученных результатов можно сделать вывод, что предлагаемая модель прогнозирования, использующая модель распределённых вычислений MapReduce, обладает хорошей масштабируемостью, что позволяет эффективно использовать данную модель при обработке большого объёма данных.

Заключение

В работе представлена распределённая пространственно-временная модель краткосрочного прогнозирования транспортного потока, основанная на методе k ближайших соседей. Описание пространства признаков в данной модели формируется на основе пространственных и временных характеристик транспортных потоков в пространственно компактной области транспортной сети.

Для распределённой обработки большого объёма информации была использована модель MapReduce, реализованная в фрэймворке с открытым исходным кодом Apache Spark. Экспериментальные исследования, проведённые по данным движения транспорта в г. Самаре, показали, что представленная модель обладает высокой точностью прогнозирования и временем работы, достаточным для прогнозирования в режиме реального времени.

Дальнейшие исследования могут быть направлены на детальное сравнение представленной модели с другими известными алгоритмами, включая алгоритмы на основе нейронных сетей и регрессии методом опорных векторов. Другим направлением дальнейших исследований является проведение экспериментальных исследований для большего числа временных интервалов.

Благодарности

Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ (уникальный идентификатор проекта RFMEFI57518X0177).

Литература

1. **Lana, I.** Road traffic forecasting: Recent advances and new challenges / I. Lana, J. Del Ser, M. Velez, E. Vlahogianni // IEEE Intelligent Transportation Systems Magazine. – 2018. – Vol. 10, Issue 2. – P. 93-109. – DOI: 10.1109/MITS.2018.2806634.
2. **Smith, B.L.** Comparison of parametric and nonparametric models for traffic flow forecasting / B.L. Smith, B.M. Williams, R.K. Oswald // Transportation Research Part C: Emerging Technologies. – 2002. – Vol. 10, Issue 4. – P. 303-321. – DOI: 10.1016/S0968-090X(02)00009-8.
3. **Smith, B.L.** Traffic flow forecasting: comparison of modeling approaches / B.L. Smith, M.J. Demetsky // Journal of Transportation Engineering –1997. – Vol. 123, Issue 4. – P. 261-266. – DOI: 10.1061/(ASCE)0733-947X(1997)123:4(261).
4. **Xia, D.** A distributed spatial-temporal weighted model on MapReduce for short-term traffic flow forecasting / D. Xia, B. Wang, H. Li, Y. Li, Z. Zhang // Neurocomputing. – 2016. – Vol. 179. – P. 246-263. – DOI: 10.1016/j.neucom.2015.12.013.
5. **Lv, Y.** Traffic flow prediction with big data: a deep learning approach / Y. Lv, Y. Duan, W. Kang, Z. Li, F.-Y. Wang // IEEE Transactions on Intelligent Transportation Systems – 2015. – Vol. 16, Issue 2. – P. 865-873. – DOI: 10.1109/TITS.2014.2345663.
6. **Vlahogianni, E.** Short-term traffic forecasting: Overview of objectives and methods / E. Vlahogianni, J. Golias, M. Karlaftis // Transport Reviews. – 2004. – Vol. 24, Issue 5. – P. 533-557. – DOI: 10.1080/0144164042000195072.
7. **Vlahogianni, E.** Short-term traffic forecasting: Where we are and where we're going / E. Vlahogianni, M. Karlaftis, J. Golias // Transportation Research Part C: Emerging Technologies. – 2014. – Vol. 43, Issue 1. – P. 3-19. – DOI: 10.1016/j.trc.2014.01.005.
8. **Karlaftis, M.G.** Statistical methods versus neural networks in transportation research: Differences, similarities and some insights / M.G. Karlaftis, E.I. Vlahogianni // Transportation Research Part C: Emerging Technologies. – 2011. – Vol. 19, Issue 3. – P. 387-389. – DOI: 10.1016/j.trc.2010.10.004.
9. **Shekhar, S.** Adaptive seasonal time series models for forecasting short-term traffic flow / S. Shekhar, B.M. Williams // Transportation Research Record. – 2007. – Vol. 2024, Issue 1. – P. 116-125. – DOI: 10.3141/2024-14.
10. **Chandra, S.R.** Predictions of freeway traffic speeds and volumes using vector autoregressive models / S.R. Chandra, H. Al-Deek // Journal of Intelligent Transportation Systems: Technology, Planning, and Operations. –2009. – Vol. 13, Issue 2. – P. 53-72. – DOI: 10.1080/15472450902858368.
11. **Guo, J.** Real-time short-term traffic speed level forecasting and uncertainty quantification using layered Kalman filters / J. Guo, B.M. Williams// Transportation Research Record. – 2010. – Vol. 2175, Issue 1. – P. 28-37. – DOI: 10.3141/2175-04.
12. **Wang, Y.** Real-time freeway traffic state estimation based on extended Kalman filter: a general approach / Y. Wang, M. Papageorgiou // Transportation Research Part B: Methodological. – 2005. – Vol. 39, Issue 2. – P. 141-167. – DOI: 10.1016/j.trb.2004.03.003.
13. **Fusco, G.** Short-term traffic predictions on large urban traffic networks: Applications of network-based machine learning models and dynamic traffic assignment models / G. Fusco, C. Colombaroni, L. Comelli, N. Isaenko // Proceeding of the Fourth IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS). – 2015. – P. 93-101. – DOI: 10.1109/MTITS.2015.7223242.
14. **Yin, H.** Urban traffic flow prediction using a fuzzy-neural approach / H. Yin, S. Wong, J. Xu, C. Wong // Transportation Research Part C: Emerging Technologies. – 2002. – Vol. 10, Issue 2. – P. 85-98. – DOI: 10.1016/S0968-090X(01)00004-3.
15. **Lana, I.** Joint feature selection and parameter tuning for short-term traffic flow forecasting based on heuristically optimized multi-layer neural networks / I. Lana, J. Del Ser, M. Velez, I. Oregi // Advances in Intelligent Systems and Computing. – 2017. – Vol. 514. – P. 91-100. – DOI: 10.1007/978-981-10-3728-3_10.
16. **Zheng, Z.** Short-term traffic volume forecasting: a k-nearest neighbor approach enhanced by constrained linearly sewing principle component algorithm / Z. Zheng, D. Su // Transportation Research Part C: Emerging Technologies. – 2014. – Vol. 43, Issue 1. – P. 143-157. – DOI: 10.1016/j.trc.2014.02.009.
17. **Cai, P.** A spatiotemporal correlative K-nearest neighbor model for short-term traffic multistep forecasting / P. Cai, Y. Wang, G. Lu, P. Chen, C. Ding, J. Sun // Transportation Research Part C: Emerging Technologies. – 2016. – Vol. 62. – P. 21-34. – DOI: 10.1016/j.trc.2015.11.002.
18. **Wu, C.-H.** Travel-time prediction with support vector regression / C.-H. Wu, J.-M. Ho, D.T. Lee // IEEE Transactions on Intelligent Transportation Systems. – 2004. – Vol. 5, Issue 4. – P. 276-281. – DOI: 10.1109/TITS.2004.837813.
19. **Su, H.** Short-term traffic flow prediction based on incremental support vector regression / H. Su, L. Zhang, S. Yu // Third

International Conference on Natural Computation. –2007. – Vol. 1. – P. 640-645. – DOI: 10.1109/ICNC.2007.661.
20. **Fei, X.** Bayesian dynamic linear model approach for real-time short-term freeway travel time prediction / X. Fei, C.-C. Lu, K.A. Liu // Transportation Research Part C: Emerging Technologies. – 2011. – Vol. 19, Issue 6. – P. 1306-1318. – DOI: 10.1016/j.trc.2010.10.005.
21. **Zhu, Z.** Short-term traffic flow prediction with linear conditional Gaussian Bayesian network / Z. Zhu, B. Peng, C. Xiong, L. Zhang // Journal of Advanced Transportation. – 2016. – Vol. 50, Issue 6. – P. 1-13. – DOI: 10.1002/atr.1392.
22. **Sun, S.** The selective random subspace predictor for traffic flow forecasting / S. Sun, C. Zhang // IEEE Transactions on Intelligent Transportation Systems. – 2007. – Vol. 8, Issue 2. – P. 367-373. – DOI: 10.1109/TITS.2006.888603.
23. **Агафонов, А.А.** Оценка и прогнозирование параметров транспортных потоков с использованием композиции методов машинного обучения и моделей прогнозирования временных рядов / А.А. Агафонов, В.В. Мясников // Компьютерная оптика. – 2014. – Т. 38, № 3. – С. 539-549. – DOI: 10.18287/0134-2452-2014-38-3-539-549.
24. **Zhang, N.** Seasonal autoregressive integrated moving average and support vector machine models: prediction of short term traffic flow on freeways / N. Zhang, Y. Zhang, H. Lu // Transportation Research Record. – 2011. – Vol. 2215, Issue 1. – P. 85-92. – DOI: 10.3141/2215-09.
25. **Moretti, F.** Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling / F. Moretti, S. Pizzuti, S. Panzieri, M. Annunziato // Neurocomputing. – 2015. – Vol. 167, Issue C. – P. 3-7. – DOI: 10.1016/j.neucom.2014.08.100.
26. **Chrobok, R.** Different methods of traffic forecast based on real data / R. Chrobok, O. Kaumann, J. Wahle, M. Schreckenberg // European Journal of Operational Research. – 2004. – Vol. 155, Issue 3. – P. 558-568. – DOI: 10.1016/j.ejor.2003.08.005.
27. **Laña, I.** Understanding daily mobility patterns in urban road networks using traffic flow analytics / I. Laña, J. Del Ser, I. Olabarrieta // Proceedings of the IEEE/IFIP Network Operations and Management Symposium. – 2016. – P. 1157-1162. – DOI: 10.1109/NOMS.2016.7502980.
28. **Jain, A.K.** Data clustering: 50 years beyond k-means / A.K. Jain // Pattern Recognition Letters. – 2010. – Vol. 31, Issue 8. – P. 651-666. – DOI: 10.1016/j.patrec.2009.09.011.
29. **Han, J.** Data mining: Concepts and techniques / J. Han, M. Kamber, J. Pei. – San Francisco: Morgan Kaufmann Publishers Inc., 2006. – 770 p. – ISBN: 978-1-55860-901-3.
30. **Myasnikov, E.V.** Hyperspectral image segmentation using dimensionality reduction and classical segmentation approaches / E.V. Myasnikov // Computer Optics. – 2017. – Vol. 41(4). – P. 564-572. – DOI: 10.18287/2412-6179-2017-41-4-564-572.
31. **Carmel, D.** Enhancing cluster labeling using wikipedia / D. Carmel, H. Roitman, N. Zwerdling // Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. – 2009. – P. 139-146. – DOI: 10.1145/1571941.1571967.
32. **Ji, Y.** On the spatial partitioning of urban transportation networks / Y. Ji, N. Geroliminis // Transportation Research Part B: Methodological. – 2012. – Vol. 46, Issue 10. – P. 1639-1656. – DOI: 10.1016/j.trb.2012.08.005.
33. **Ding, C.H.Q.** A min-max cut algorithm for graph partitioning and data clustering / C.H.Q. Ding, X. He, H. Zha, M. Gu, H.D. Simon // Proceedings of the 2001 IEEE International Conference on Data Mining. – 2001. – P. 107-114. – DOI: 10.1109/ICDM.2001.989507.
34. Apache Spark™ [Electronical Resource]. – URL: https://spark.apache.org/ (request date 28.08.2018).
35. **Dean, J.** MapReduce: simplified data processing on large clusters / J. Dean, S. Ghemawat // Communications of the ACM. – 2008. – Vol. 51, Issue 1. – P. 107-113. – DOI: 10.1145/1327452.1327492.

Сведения об авторах

Агафонов Антон Александрович, 1988 года рождения. В 2011 году окончил Самарский государственный аэрокосмический университет (СГАУ), в 2014 защитил диссертацию на соискание степени кандидата технических наук. В настоящее время работает старшим научным сотрудником НИЛ-55 Самарского университета. Круг научных интересов включает геоинформационные технологии, транспортное моделирование, веб-технологии. Имеет 12 публикаций, из них 5 статей. E-mail: *ant.agafonov@gmail.com*.

Юмаганов Александр Сергеевич, 1993 года рождения. В 2016 году окончил Самарский национальный исследовательский университет имени академика С.П. Королева (Самарский университет) по специальности «Информационная безопасность автоматизированных систем». В настоящее время является аспирантом Самарского университета. Область научных интересов: программирование, распознавание образов. Имеет 6 публикаций. E-mail: *yumagan@gmail.com*.

Сведения об авторе **Мясников Владислав Валерьевич,** см. стр.1006 этого номера.

ГРНТИ: 20.23.27
Поступила в редакцию 3 декабря 2018 г. Окончательный вариант – 10 декабря 2018 г.

BIG DATA ANALYSIS IN A GEOINFORMATIC PROBLEM OF SHORT-TERM TRAFFIC FLOW FORECASTING BASED ON A *K* NEAREST NEIGHBORS METHOD

A.A. Agafonov [1], A.S. Yumaganov [1], V.V. Myasnikov [1,2]
[1] Samara National Research University, 443086, Russia, Samara, Moskovskoye Shosse 34,
[2]IPSI RAS – Branch of the FSRC "Crystallography and Photonics" RAS, Molodogvardeyskaya 151, 443001, Samara, Russia

Abstract

Accurate and timely information on the current and predicted traffic flows is important for the successful deployment of intelligent transport systems. These data play an essential role in traffic

management and control. Using traffic flow information, travelers could plan their routes to avoid traffic congestion, reduce travel time and environmental pollution, as well as improving traffic operation efficiency in general. In this paper, we propose a distributed model for short-term traffic flow prediction based on a *k* nearest neighbors method, that takes into account spatial and temporal traffic flow distributions. The proposed model is implemented as a MapReduce based algorithm in an Apache Spark framework. An experimental study of the proposed model is carried out on a traffic flow data in the transportation network of Samara, Russia. The results demonstrate that the proposed model has high predictive accuracy and an execution time sufficient for real-time prediction.

Keywords: traffic flow, short-term forecasting, *k* nearest neighbors, MapReduce.

Citation: Agafonov AA, Yumaganov AS, Myasnikov VV. Big data analysis in a geoinformatic problem of short-term traffic flow forecasting based on a *k* nearest neighbors method. Computer Optics 2018; 42(6): 1101-1111. DOI: 10.18287/2412-6179-2018-42-6-1101-1111.

Acknowledgements: The work was supported by the Ministry of Science and Higher Education of the Russian Federation (unique project identifier RFMEFI57518X0177).

References

[1] Lana I, Del Ser J, Velez M. Vlahogianni E. Road traffic forecasting: Recent advances and new challenges. IEEE Intelligent Transportation Systems Magazine 2018; 10(2): 93-109. DOI: 10.1109/MITS.2018.2806634.

[2] Smith BL, Williams BM. Oswald RK. Comparison of parametric and nonparametric models for traffic flow forecasting. Transportation Research Part C: Emerging Technologies 2002; 10(4): 303-321. DOI: 10.1016/S0968-090X(02)00009-8

[3] Smith BL, Demetsky MJ. Traffic flow forecasting: comparison of modeling approaches. Journal of Transportation Engineering 1997; 123(4): 261-266. DOI: 10.1061/(ASCE)0733-947X(1997)123:4(261)

[4] Xia D, Wang B, Li H, Li Y, Zhang Z. A distributed spatial-temporal weighted model on MapReduce for short-term traffic flow forecasting. Neurocomputing 2016; 179: 246-263. DOI: 10.1016/j.neucom.2015.12.013.

[5] Lv Y, Duan Y, Kang W, Li Z, Wang FY. Traffic flow prediction with big data: a deep learning approach. IEEE Transactions on Intelligent Transportation Systems 2015; 16(2): 865-873. DOI: 10.1109/TITS.2014.2345663.

[6] Vlahogianni E, Golias J, Karlaftis M. Short-term traffic forecasting: Overview of objectives and methods. Transport Reviews 2004; 24(5): 533-557. DOI: 10.1080/0144164042000195072.

[7] Vlahogianni E, Karlaftis M, Golias J. Short-term traffic forecasting: Where we are and where we're going. Transportation Research Part C: Emerging Technologies 2014; 43(1): 3-19. DOI: 10.1016/j.trc.2014.01.005.

[8] Karlaftis MG, Vlahogianni EI. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. Transportation Research Part C: Emerging Technologies 2011; 19(3): 387-389. DOI: 10.1016/j.trc.2010.10.004.

[9] Shekhar S, Williams BM. Adaptive seasonal time series models for forecasting short-term traffic flow. Transportation Research Record 2007; 2024(1): 116-125. DOI: 10.3141/2024-14.

[10] Chandra SR, Al-Deek H. Predictions of freeway traffic speeds and volumes using vector autoregressive models. Journal of Intelligent Transportation Systems: Technology, Planning, and Operations 2009; 13(2): 53-72. DOI: 10.1080/15472450902858368.

[11] Guo J, Williams BM. Real-time short-term traffic speed level forecasting and uncertainty quantification using layered Kalman filters. Transportation Research Record 2010; 2175(1): 28-37. DOI: 10.3141/2175-04.

[12] Wang Y, Papageorgiou M. Real-time freeway traffic state estimation based on extended Kalman filter: a general approach. Transportation Research Part B: Methodological 2005; 39(2); 141-167. DOI: 10.1016/j.trb.2004.03.003.

[13] Fusco G, Colombaroni C, Comelli L, Isaenko N. Short-term traffic predictions on large urban traffic networks: Applications of network-based machine learning models and dynamic traffic assignment models. Proc 4th IEEE Int Conf Models and Technologies for Intelligent Transportation Systems (MT-ITS) 2015: 93-101. DOI: 10.1109/MTITS.2015.7223242.

[14] Yin H, Wong S, Xu J, Wong C. Urban traffic flow prediction using a fuzzy-neural approach. Transportation Research Part C: Emerging Technologies 2002; 10(2): 85-98. DOI: 10.1016/S0968-090X(01)00004-3.

[15] Lana I, Del Ser J, Velez M, Oregi I. Joint feature selection and parameter tuning for short-term traffic flow forecasting based on heuristically optimized multi-layer neural networks. Advances in Intelligent Systems and Computing 2017; 514: 91-100. DOI: 10.1007/978-981-10-3728-3_10.

[16] Zheng Z, Su D. Short-term traffic volume forecasting: a k-nearest neighbor approach enhanced by constrained linearly sewing principle component algorithm. Transportation Research Part C: Emerging Technologies 2014; 43(1): 143-157. DOI: 10.1016/j.trc.2014.02.009.

[17] Cai P, Wang Y, Lu G, Chen P, Ding C, Sun J. A spatiotemporal correlative K-nearest neighbor model for short-term traffic multistep forecasting. Transportation Research Part C: Emerging Technologies 2016; 62: 21-34. DOI: 10.1016/j.trc.2015.11.002.

[18] Wu CH, Ho JM, Lee DT. Travel-time prediction with support vector regression. IEEE Transactions on Intelligent Transportation Systems 2004; 5(4): 276-281. DOI: 10.1109/TITS.2004.837813.

[19] Su H, Zhang L, Yu S. Short-term traffic flow prediction based on incremental support vector regression. Third International Conference on Natural Computation 2007; 1: 640-645. DOI: 10.1109/ICNC.2007.661.

[20] Fei X, Lu CC, Liu KA. Bayesian dynamic linear model approach for real-time short-term freeway travel time prediction. Transportation Research Part C: Emerging Technologies 2011; 19(6): 1306-1318. DOI: 10.1016/j.trc.2010.10.005.

[21] Zhu Z, Peng B, Xiong C, Zhang L. Short-term traffic flow prediction with linear conditional Gaussian Bayesian network. Journal of Advanced Transportation 2016; 50(6): 1-13. DOI: 10.1002/atr.1392.

[22] Sun S, Zhang C. The selective random subspace predictor for traffic flow forecasting. IEEE Transactions on Intelli-

gent Transportation Systems 2007; 8(2): 367-373. DOI: 10.1109/TITS.2006.888603.
[23] Agafonov A, Myasnikov V. Traffic flow forecasting algorithm based on combination of adaptive elementary predictors. In Book: Khachay MYu, Konstantinova N, Panchenko A, Ignatov D, Labunets VG, eds. Analysis of Images, Social Networks and Texts: AIST 2015. Cham: Springer; 2015: 163-174. DOI: 10.1007/978-3-319-26123-2_16.
[24] Zhang N, Zhang Y, Lu H. Seasonal autoregressive integrated moving average and support vector machine models: prediction of short term traffic flow on freeways. Transportation Research Record 2011; 2215(1): 85-92. DOI: 10.3141/2215-09
[25] Moretti F, Pizzuti S, Panzieri S, Annunziato M. Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling. Neurocomputing 2015; 167(C): 3-7. DOI: 10.1016/j.neucom.2014.08.100.
[26] Chrobok R, Kaumann O, Wahle J, Schreckenberg M. Different methods of traffic forecast based on real data. European Journal of Operational Research 2004; 155(3): 558-568. DOI: 10.1016/j.ejor.2003.08.005.
[27] Laña I, Del Ser J, Olabarrieta I. Understanding daily mobility patterns in urban road networks using traffic flow analytics. Proc IEEE/IFIP Network Operations and Management Symposium 2016: 1157-1162. DOI: 10.1109/NOMS.2016.7502980.
[28] Jain AK. Data clustering: 50 years beyond k-means. Pattern Recognition Letters 2010; 31(8): 651-666. DOI: 10.1016/j.patrec.2009.09.011.
[29] Han J, Kamber M, Pei J. Data mining: Concepts and techniques. San Francisco: Morgan Kaufmann Publishers Inc; 2006. ISBN: 978-1-55860-901-3.
[30] Myasnikov EV. Hyperspectral image segmentation using dimensionality reduction and classical segmentation approaches. Computer Optics 2017; 41(4): 564-572. DOI: 10.18287/2412-6179-2017-41-4-564-572.
[31] Carmel D, Roitman H, Zwerdling N. Enhancing cluster labeling using wikipedia. Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval 2009: 139-146. DOI: 10.1145/1571941.1571967.
[32] Ji Y, Geroliminis N. On the spatial partitioning of urban transportation networks. Transportation Research Part B: Methodological 2012; 46(10): 1639-1656. DOI: 10.1016/j.trb.2012.08.005.
[33] Ding CHQ, He X, Zha H, Gu M, Simon HD. A min-max cut algorithm for graph partitioning and data clustering. Proc 2001 IEEE International Conference on Data Mining 2001: 107-114. DOI: 10.1109/ICDM.2001.989507.
[34] Apache Spark™. Source: ⟨https://spark.apache.org/⟩.
[35] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Communications of the ACM 2008; 51(1): 107-113. DOI: 10.1145/1327452.1327492.

Authors' information

Anton Aleksandrovich Agafonov (b. 1988) graduated from Samara State Aerospace University (SSAU) at 2011, received his PhD in Techical Sciences at 2014. At present, he is a researcher at Samara University. The area of interests includes geoinformatics, transport modelling and web-technologies. He's list of publications contains 12 publications, including 5 scientific papers. E-mail: *ant.agafonov@gmail.ru* .

Alexander Sergeevich Yumaganov (b. 1993) graduated from Samara National Research University (2016), programme – Information Security of Automated Systems. Nowadays he is postgraduate at Samara National Research University. His interests include computer programming, pattern recognition. He has 6 publications. E-mail: *yumagan@gmail.com* .

The information about author **Vladislav Valerievich Myasnikov** you can find on page 1007 of this issue.

Received December 3, 2018. The final version – December 10, 2018.