Комплексирование данных из разнородных источников в задачах моделирования транспортных потоков*

В. М. Старожилец^{1,2}, Ю. В. Чехович^{1,2}

starvsevol@gmail.com; chehovich@forecsys.ru

¹ФИЦ «Информатика и управление» РАН, Россия, г. Москва, ул. Вавилова, д. 44/2

²Московский физико-технический институт, Россия, г. Долгопрудный, Институтский пер., д. 9

Исследуется задача агрегации данных с GPS-треков и дорожных датчиков для построения и решения разностной схемы, соответствующей выбранной математической модели транспортного потока. Отдельно рассматриваются ситуации транспортного потока на самой автомагистрали и потока на въездах и съездах. Для решения обеих задач предложены алгоритмы, а также проведены эксперименты на реальных данных с использованием этих алгоритмов. Для проведения вычислительных экспериментов использованы анонимные трековые данные от сервиса Яндекс.Пробки и данные с дорожных датчиков Центра организации дорожного движения. В качестве автомагистрали рассматривалась Московская кольцевая автомобильная дорога.

Ключевые слова: восстановление данных; МКАД; данные GPS-треков; дорожные датчики

DOI: 10.21469/22233792.2.3.01

1 Введение

Работа посвящена проблеме агрегации данных из разных источников, используемых в задаче моделирования транспортных потоков [1]. Для моделирования транспортных потоков используются данные о скорости и числе проехавших по рассматриваемому участку автодороги автотранспортных средств (АТС). Эти данные могут быть получены с помощью *GPS-треков* (данные трекового типа) и *дорожных датчиков*, которые имеют различные свойства. Дорожные датчики измеряют скорость и число АТС с приемлемой точностью, погрешность измерений не превышает 10% в зависимости от типа датчика [2], но не всегда покрывают транспортную сеть на требуемом уровне. В то же время данные с GPS-треков (трекового типа) имеют недостаточную точность, так как фиксируют небольшой процент¹ от общего числа проехавших АТС, однако полностью покрывают транспортную сеть. Поэтому основная идея работы заключается в том, чтобы агрегировать данные с дорожных датчиков и GPS-треков для получения более точных данных с большим покрытием транспортной сети.

Моделирование транспортных потоков основано на их сходстве с жидкой или газовой средой. В частности, базовая модель Лайтхилла–Уизема–Ричардса (LWR) [3–5] основана на предположении о существовании взаимно-однозначной зависимости между скоростью и плотностью потока АТС и сохранении числа АТС в транспортной сети. В современном макроскопическом подходе транспортный поток описывается нелинейной системой

*Работа выполнена при частичной финансовой поддержке РФФИ проект № 14-07-00685

¹Рассчитать долю трековых АТС в общем потоке можно, взяв отношение числа трековых АТС, к числу АТС зафиксированному датчиком для детектора и сегмента под ним. Рассчитанные по данной методике на имеющихся у авторов данных значения имеют большой разброс.

гиперболических дифференциальных уравнений в частных производных второго порядка в различных постановках [6–13]. Такая система требует точных входных данных с хорошим покрытием транспортной сети для выбранной модели. Эти требования к данным необходимы для построения и разрешения разностной схемы, соответствующей выбранной модели, и заданию граничных условий. Ранее было показано [14], что абсолютно точных данных без пропусков для задач транспортного типа не существует.

Ранее задача агрегирования произвольных данных рассматривалась в работах [15–18]. Предложенные там методы не учитывают специфику задачи агрегации данных для моделирования транспортных потоков. Для решения проблемы получения точных данных с хорошим покрытием транспортной сети в работе предлагается метод агрегирования данных GPS-треков и дорожных датчиков. Однако GPS-треки на автомагистрали и на въездах и съездах с нее сильно отличаются из-за существенно меньшего потока АТС на въездах и съездах по сравнению с автомагистралью, поэтому данные с въездов и съездов рассматриваются отдельно от данных с автомагистрали. В частности, для данных с автомагистрали предложен метод агрегации данных GPS-треков и дорожных датчиков, основанный на построении линейной модели для скорости и числа АТС. Критерием качества полученной модели является среднеквадратичная ошибка между оцененным числом проехавших АТС и реальным, а также коэффициент корреляции между ними. Число реально проехавших АТС определяется по данным дорожных датчиков. Для данных на въездах и съездах был разработан метод восстановления суммарного потока на основе сохранения числа АТС в транспортной сети с использованием оценки числа проехавших АТС, полученной с помощью модели на данных с автомагистрали. В работе был проведен вычислительный эксперимент на данных дорожных датчиков и GPS-треков для Московской кольцевой автомобильной дороги за 2012 г.

2 Постановка задачи

Поскольку в работе рассматривается задача агрегации данных для двух принципиально различных дорожных конфигураций: автомагистрали и въезд–съезд, то необходимо поставить задачу для каждой из этих конфигураций. Поставленные задачи будут отличаться подходом к агрегации данных и способом проверки полученного решения.

2.1 Задача агрегации для данных с автомагистрали

Пусть $N_{\text{track}} \in \mathbb{N}$ и $V_{\text{track}} \in \mathbb{R}_+$ — число АТС и их средняя скорость, полученные из данных GPS-треков для определенного участка дороги. Обозначим через $N_{\text{det}} \in \mathbb{N}$ и $V_{\text{det}} \in \mathbb{R}_+$ число АТС и их среднюю скорость, полученные с помощью дорожных датчиков для того же участка дороги.

Необходимо найти функцию $f : \mathbb{N} \times \mathbb{R}_+ \times \mathbb{R}^m \rightarrow \mathbb{R}_+$ от $N_{\text{track}} \times V_{\text{track}}$ и вектора параметров $\mathbf{a} \in \mathbb{R}^m$, где m — сложность модели, которая аппроксимирует число АТС, проехавших по сегменту автомагистрали

$$N_{\text{est}} = f(\mathbf{a} | N_{\text{track},i}, V_{\text{track},i})$$

и является решением следующей задачи:

$$\sigma(\mathbf{a} | \mathbf{N}_{\text{track}}, \mathbf{V}_{\text{track}}, \mathbf{N}_{\text{det}}) = \sqrt{\frac{1}{n} \sum_{i=1}^n (f(\mathbf{a} | N_{\text{track},i}, V_{\text{track},i}) - N_{\text{det},i})^2} \rightarrow \min_{\mathbf{a}}, \quad (1)$$

где $\mathbf{N}_{\text{track}} = [N_{\text{track},i}] \in \mathbb{N}^n$, $\mathbf{V}_{\text{track}} = [V_{\text{track},i}] \in \mathbb{R}_+^n$ и $\mathbf{N}_{\text{det}} = [N_{\text{det},i}] \in \mathbb{N}^n$ — вектора значений N_{track} , V_{track} и N_{det} в момент времени i , а n — число двухминутных интервалов

в выбранном временном промежутке. Предполагается, что f зависит только от трековых переменных $\mathbf{N}_{\text{track}}$ и $\mathbf{V}_{\text{track}}$, что позволяет ее использовать на участках автомагистрали, которые не покрываются дорожными датчиками.

Однако скорость V_{track} вычисляется по малой доле АТС, поэтому она может сильно отличаться от реальной скорости потока АТС. Для уменьшения влияния этого отличия предлагается вместо оригинальных значений V_{track} использовать модифицированные значения V_{est} , определяемые выражением:

$$V_{\text{est}} = b_1 + b_2 V_{\text{track}},$$

где коэффициенты b_1 и b_2 являются решением следующей задачи:

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (b_1 + b_2 V_{\text{track},i} - V_{\text{det},i})^2} \rightarrow \min_{b_1, b_2}, \quad (2)$$

а $V_{\text{track},i} \in \mathbb{R}_+$ и $V_{\text{det},i} \in \mathbb{R}_+$ — значения V_{track} и V_{det} в момент времени i . Таким образом, задача (1) преобразуется в

$$\sigma(\mathbf{a} | \mathbf{N}_{\text{track}}, \mathbf{V}_{\text{est}}, \mathbf{N}_{\text{det}}) \rightarrow \min_{\mathbf{a}}. \quad (3)$$

Задача (3) решается на сегментах, для которых известны как данные с GPS-треков, так и данные с дорожных датчиков. Это множество сегментов разбивается на обучающую и контрольные выборки. Способ разбиения изложен в подразд. 5.2.

Внешним критерием качества полученного решения является коэффициент корреляции $\text{corr}(\mathbf{N}_{\text{est}}, \mathbf{N}_{\text{det}})$, где $\mathbf{N}_{\text{est}} = [N_{\text{est},i}] \in \mathbb{R}_+^n$ и $\mathbf{N}_{\text{det}} = [N_{\text{det},i}] \in \mathbb{N}^n$ — вектора значений N_{est} и N_{det} в момент времени i .

2.2 Задача агрегации для данных со въездов-съездов

Въезды и съезды образуют перекрестки с автомагистралью. Перекресток — это место пересечения, примыкания или разветвления дорог на одном уровне, ограниченное воображаемыми линиями, соединяющими соответственно противоположные, наиболее удаленные от центра перекрестка начала закруглений проезжих частей.

Для оценки суммарного количества АТС, въехавших и съехавших с автомагистрали, используется уравнение баланса, заключающееся в том, что на въездах и съездах число въезжающих АТС должно равняться числу выезжающих:

$$N_{\text{ain}} + N_{\text{in}} = N_{\text{aout}} + N_{\text{out}},$$

где $N_{\text{ain}} \in \mathbb{R}_+$ и $N_{\text{aout}} \in \mathbb{R}_+$ — оценка числа въехавших на перекресток по автомагистрали и выехавших по автомагистрали после перекрестка АТС, полученная после решения задачи (3); N_{in} — суммарное число въехавших по въездам АТС; N_{out} — суммарное число съехавших по съездам АТС.

Рассмотрим, как вычислить значение N_{in} ; значение N_{out} вычисляется аналогично. По определению $N_{\text{in}} = \sum_{k \in K_{\text{in}}} N_{\text{det},k}$, где $K_{\text{in}} = \{1, \dots, K\}$ — множество индексов съездов, а $N_{\text{det},k}$ — значение N_{det} на k -м въезде. Но не для всех $k \in K_{\text{in}}$ значение $N_{\text{det},k}$ известно. Поэтому рассмотрим разбиение множества K_{in} на два непересекающихся подмножества K_{indet} и K_{intrack} : $K_{\text{in}} = K_{\text{indet}} \cup K_{\text{intrack}}$ и $K_{\text{intrack}} \cap K_{\text{indet}} = \emptyset$. Множество K_{indet} состоит из индексов въездов, для которых известно N_{det} , а множество K_{intrack} состоит из индексов въездов, для которых неизвестно N_{det} .

Чтобы получить оценку N_{det} для въездов из K_{intrack} , предлагается использовать подход описанный в подразд. 2.1:

$$\sigma(\mathbf{a}|\mathbf{N}_{\text{track}}, \mathbf{V}_{\text{est}}, \mathbf{N}_{\text{det}}) \rightarrow \min_{\mathbf{a}},$$

где $\mathbf{N}_{\text{track}} = [N_{\text{track},i}] \in \mathbb{N}^n$, $\mathbf{V}_{\text{track}} = [V_{\text{track},i}] \in \mathbb{R}_+^n$ и $\mathbf{N}_{\text{det}} = [N_{\text{det},i}] \in \mathbb{N}^n$ — вектора значений N_{track} , V_{track} и N_{det} для въезда из множества K_{indet} в момент времени $i \in I_{\text{in}}$, $n = |I_{\text{in}}|$. Множество I_{in} — это множество индексов временных интервалов, таких что в момент времени $i \in I_{\text{in}}$ известно число въехавших АТС $N_{\text{in},k}$ для всех въездов $k \in K_{\text{in}}$. Для получение необходимых данных на съездах нужно решить аналогичную задачу. Введем суммарное число въехавших по въездам АТС

$$N_{\text{in}} = \sum_{k \in K_{\text{indet}}} N_{\text{det},k} + \sum_{k' \in K_{\text{intrack}}} N_{\text{est},k'},$$

где $N_{\text{det},k}$ — значение N_{det} на k -м въезде; $N_{\text{est},k'}$ — значение N_{est} на k' -м въезде. Аналогично определяется суммарное число съехавших по съездам АТС N_{out} .

Требуется построить алгоритм нахождения таких значений N_{estin} и N_{estout} , чтобы они удовлетворяли уравнению баланса и различие между ними и значениями N_{in} и N_{out} в моменты времени из множества I_{in} и I_{out} соответственно было не слишком велико, что формализуется следующим образом:

$$\left. \begin{aligned} & (N_{\text{ain}} + N_{\text{estin}} - N_{\text{aout}} - N_{\text{estout}})^2 \rightarrow \min_{N_{\text{estin}}, N_{\text{estout}}}; \\ \text{s.t. } & \frac{1}{n} \left(\sum_{i \in I_{\text{in}}} |N_{\text{estin}}^i - N_{\text{in}}^i| + \sum_{i' \in I_{\text{out}}} |N_{\text{estout}}^{i'} - N_{\text{out}}^{i'}| \right) < \delta, \end{aligned} \right\} \quad (4)$$

где δ — допустимое отличие оценки числа АТС на съездах и въездах $N_{\text{estout}}^{i'}$ и N_{estin}^i от наблюдения $N_{\text{out}}^{i'}$ и N_{in}^i ; I_{out} — множество аналогичное I_{in} для съездов; N_{in}^i и N_{estin}^i — значения N_{in} и N_{estin} в момент времени i ; $N_{\text{out}}^{i'}$ и $N_{\text{estout}}^{i'}$ — значения N_{out} и N_{estout} в момент времени i' ; $n = |I_{\text{in}}| + |I_{\text{out}}|$.

3 Выбор модели для предсказания числа автотранспортных средств

Для решения задачи (3) необходимо задать вид функции f . Было показано, что f линейно зависит от числа трековых АТС N_{track} , скорости V_{track} и оценки плотности потока [19]. Для проверки наличия дополнительных зависимостей была построена скрипичная диаграмма [20], изображенная на рис. 1. Рисунок 1 показывает, что линейной аппроксимации недостаточно для приближения реального числа АТС. Чтобы учесть найденную нелинейность, предлагается добавить слагаемое $\log(N_{\text{track}})$, так как средние значения на рис. 1 лежат на кривой, похожей на график логарифма. Таким образом, будем искать функцию f в виде:

$$f(\mathbf{a}|N_{\text{track},i}, V_{\text{est},i}) = a_0 + a_1 N_{\text{track},i} + a_2 \log(N_{\text{track},i}) + a_3 V_{\text{est},i} + a_4 \frac{N_{\text{track},i}}{V_{\text{est},i}}. \quad (5)$$

Также кроме модели вида (5) были рассмотрены и другие модели, такие как модель числа трековых АТС:

$$f(\mathbf{a}|N_{\text{track},i}) = a_0 + a_1 N_{\text{track},i}, \quad (6)$$

Рис. 1 Скрипичная диаграмма зависимости полного числа АТС от числа тревых АТС

модель числа АТС и скорости

$$f(\mathbf{a}|N_{\text{track},i}, V_{\text{est},i}) = a_0 + a_1 N_{\text{track},i} + a_2 V_{\text{est},i} \quad (7)$$

и модель с логарифмом числа АТС и скоростью

$$f(\mathbf{a}|N_{\text{track},i}, V_{\text{est},i}) = a_0 + a_1 \log(N_{\text{track},i}) + a_2 V_{\text{est},i}. \quad (8)$$

4 Алгоритм восстановления суммарного потока на въездах и съездах

Алгоритм восстановления суммарного потока на въездах и съездах состоит из следующих шагов:

- задать перекресток, на котором будет решаться задача восстановления суммарного потока, и определить сегменты, соответствующие въездам и съездам, а также сегменты с которых берутся данные о N_{ain} и N_{aout} ;
- определить въезды и съезды, принадлежащие множествам K_{intrack} и K_{outtrack} . Уже имеющееся на них небольшое число данных используем для определения параметров случайного Пуассоновского процесса, соответствующего данному въезду или съезду;
- используя полученные распределения на въездах и съездах из множеств K_{intrack} и K_{outtrack} вместо реальных данных, а также данные с самого МКАД (которые считаем верно восстановленными в подразд. 5.2) и данные с въездов и съездов из множеств K_{indet} и K_{outdet} , решаем задачу (4).

Алгоритм 1 Алгоритм восстановления суммарных значений потока на въездах/съездах, использующий уравнение баланса

```

если  $N \neq 0$  то
  если  $N > 0$  то
     $\max\_extra\_cars = N_{out} + (\max\_cars\_in - N_{in})$ 
    если  $N < \max\_extra\_cars$  то
       $N_{estout} = N_{out} - N \cdot N_{out} / \max\_extra\_cars$ 
       $N_{estin} = N_{in} + N \cdot (\max\_cars\_in - N_{in}) / \max\_extra\_cars$ 
    иначе
       $N_{estout} = 0$ 
       $N_{estin} = \max\_cars\_in$ 
  иначе
     $\max\_extra\_cars = N_{in} + (\max\_cars\_out - N_{out})$ 
    если  $|N| < \max\_extra\_cars$  то
       $N_{estin} = N_{in} - N \cdot N_{in} / \max\_extra\_cars$ 
       $N_{estout} = N_{out} + N \cdot (\max\_cars\_out - N_{out}) / \max\_extra\_cars$ 
    иначе
       $N_{estin} = 0$ 
       $N_{estout} = \max\_cars\_out$ 

```

Для решения поставленной задачи предлагается использовать алгоритм 1, использующий понятие дисбаланса $N = N_{ain} + N_{in} - N_{aout} - N_{out}$. Также введем следующие обозначения $\max_cars_in = k \max_cars_per_enter/exit$ — максимальное число въезжающих по въездам АТС; $\max_cars_out = k' \max_cars_per_enter/exit$ — максимальное число съезжающих по съездам АТС, где $\max_cars_per_enter/exit$ — максимальное число АТС, которые могут проехать по съезду/въезду за 2 мин; k — число въездов; k' — число съездов. Значение $\max_cars_per_enter/exit$ равно 60 АТС [21], т.е. 1 АТС за каждые 2 с.

5 Вычислительный эксперимент

В работе проведен вычислительный эксперимент на реальных данных с МКАДа за 2012 г. и проверен предложенный подход к агрегации данных с автомагистралей.

5.1 Описание данных

В данной работе используются анонимные данные с GPS-треков и дорожных датчиков за 2012 год. Данные с GPS-треков представляют собой набор сегментов, каждый из которых соотносится с некоторым участком автодороги. Объединение сегментов покрывает весь рассматриваемый участок транспортной сети. Для каждого сегмента известно число проехавших за двухминутный интервал трековых АТС N_{track} и среднее время их проезда по нему, из которого впоследствии рассчитывается скорость V_{track} . Также если за 2 мин по данному сегменту проехало менее трех трековых АТС, то они не учитываются.

Для каждого датчика известно его местоположение на автомагистрали. Данные с датчиков состоят из числа проехавших за 2 мин АТС N_{det} для каждой из полос и их скорости V_{det} . Заметим, что в данных с датчиков также могут быть ошибки, например на рис. 2 показаны данные датчика, в которых отсутствуют записи за 4 ч.

5.2 Эксперимент на автомагистрали

В экспериментах вместо числа АТС использовалась плотность АТС, чтобы учесть различные длины сегментов. Обозначим плотность АТС на участке автомагистрали, полу-

Рис. 2 Число зарегистрированных датчиком АТС в зависимости от времени суток. Детектор перестает фиксировать АТС в 17:25 и начинает в 21:15

ченную с помощью данных дорожных датчиков $\rho_{\text{det}} = N_{\text{det}}/l$, где l — длина участка автомагистрали в данных GPS-треков. $\rho_{\text{est}} \in \mathbb{R}_+^n$ и $\rho_{\text{det}} \in \mathbb{R}_+^n$ — вектора соответствующих по времени значений ρ_{est} и ρ_{det} , где n — число двухминутных интервалов в выбранном временном промежутке.

Для решения задачи (3) необходимо сначала получить преобразование скорости, решив задачу (2). Решением задачи (2) является следующее выражение:

$$V_{\text{est}} = 12,4 + 0,639V_{\text{track}}. \quad (9)$$

На рис. 3 показана зависимость плотности АТС от времени суток в случае использования преобразования (9) (рис. 3, а) и в случае использования скоростей, полученных с GPS-треков V_{track} (рис. 3, б). При использовании преобразования (9) ошибка аппроксимации на обучении $\sigma(\mathbf{a}_1|\rho_{\text{track}}, \mathbf{V}_{\text{est}}, \rho_{\text{det}}) = 0,03$ и корреляция $\text{corr}(\rho_{\text{est}}, \rho_{\text{det}}) = 0,787$, в то время как при использовании данных с GPS-треков ошибка аппроксимации $\sigma(\mathbf{a}_2|\rho_{\text{track}}, \mathbf{V}_{\text{track}}, \rho_{\text{det}}) = 0,042$ и корреляция $\text{corr}(\rho_{\text{est}}, \rho_{\text{det}}) = 0,672$. Это означает, что использование преобразования (9) повышает качество аппроксимации.

Далее рассмотрим метод улучшения качества аппроксимации с помощью построения нескольких моделей для данных с различными значениями плотности АТС, определяемыми по данным датчиков. Для этого возьмем множество данных за февраль L для датчика и подсегмента под ним и выделим из них множество $H \subset L$ — данные, соответствующие

Рис. 3 Графики полученных с помощью моделей плотностей в случае с использованием модели для скорости (а) и без ее использования (б)

плотности АТС более 0,05 АТС/м, которая является переходной между фазами свободного и синхронизированного потока для МКАД [1].

Задача (3) решается для данных L и H отдельно с получением векторов параметров $\mathbf{a}_L$ и $\mathbf{a}_H$ соответственно. Решением задачи (3) на данных H является следующее выражение:

$$N_{\text{est}} = 157,78 + 4,54N_{\text{track}} - 4,59 \log(N_{\text{track}}) + 0,153V_{\text{est}} - 85,069 \frac{N_{\text{track}}}{V_{\text{track}}}, \quad (10)$$

а на данных L :

$$N_{\text{est}} = 117,75 + 2,11N_{\text{track}} + 41,55 \log(N_{\text{track}}) - 0,327V_{\text{est}} - 128,89 \frac{N_{\text{track}}}{V_{\text{est}}}. \quad (11)$$

Первая модель использовалась при плотности выше 0,1 АТС/м (плотность, соответствующая началу синхронизированной фазы [1]), вторая — при плотности ниже 0,1 АТС/м. Корреляция на обучении составила 0,787, средняя ошибка — 0,03, а сравнение результата ρ_{det} с ρ_{est} показано на рис. 4, а. Для контроля выбраны четыре пары датчик–сегмент и проведены вычисления с использованием моделей (10) и (11). На контроле корреляция составила 0,823, 0,80, 0,85, и 0,65, средняя ошибка — 0,0363, 0,0382, 0,0339 и 0,0393 соответственно. Сравнение результата ρ_{det} с ρ_{est} показано на рис. 4, б для одной из тестовых пар.

Также был проведен эксперимент для проверки возможности использования построенной модели для получения оценки количества проехавших АТС в режиме реального времени с использованием данных GPS-треков. Чтобы обучить модель в этом случае, необходимо использовать исторические данные за некоторый промежуток времени до дня, для которого надо получить оценку. В рассматриваемом случае для обучения брались данные за каждые 7 дней перед рассматриваемым днем, который является контрольным временным интервалом. На рис. 5 показана зависимость функции ошибки и коэффициента корреляции от дня на усредненных за 10 мин данных для построенных моделях для каждого дня.

На рис. 6 показаны результаты работы моделей (6)–(8). Сравнение рассматриваемых моделей приведено в табл. 1. Из табл. 1 следует, что лучшей является модель (5).

Рис. 4 Средняя за 10 мин плотность АТС, рассчитанная по данным датчика и аппроксимированным трековым данным на обучении (а) и на контроле (б)

Рис. 5 Корреляция для усреднённых за 10 мин данных (а) и среднеквадратичная ошибка на контроле для эксперимента с обучением по 7 дням (б)

В табл. 2 приведено сравнение среднеквадратической ошибки моделей (5)–(7), (8) при большой плотности $\rho_{\text{det}} > 0,2$. Из табл. 2 следует, что модель (5) значительно лучше при больших плотностях, чем модели (6) и (8) и лучше модели (7).

5.3 Эксперимент на въездах и съездах

Приведем на примере двух перекрестков результат восстановления числа въехавших АТС за 02.02.2012. На одном из перекрестков датчиками закрыты все въезды и выезды, кроме одного въезда, результат работы алгоритма 1 представлен на рис. 7. На рис. 7 показано, что результат работы алгоритма 1 (красная кривая) проходит в области, соответствующей сумме данных с датчика и данных с GPS-треков (зеленые точки), т.е. ограничение в задаче (4) выполняется с достаточной точностью.

На втором перекрестке датчиками закрыты оба въезда, но один из датчиков фиксировал проехавшие АТС полчаса за день (до 5 АТС), и этот въезд был также включен в множество K_{intrack} , данных же трекового типа на данном въезде нет, результат работы алгоритма 1 представлен на рис. 8. На рис. 8 показано, что результат работы алгоритма 1 (красная кривая) проходит в области, соответствующей сумме данных с датчика и данных с GPS-треков (зеленые точки), т.е. ограничение в задаче (4) выполняется с достаточной точностью. Также на рис. 8 показано, что N_{estin} в большинстве двухминутных интервалов отличается от данных закрытого датчиком въезда (синяя кривая) менее чем на 5 АТС, а иногда полностью совпадает с ним. Таким образом, данные с закрытого плохим датчиком въезда подтверждаются (малое число зафиксированных АТС), а также становится понятна причина отсутствия данных трекового типа на данном въезде — слишком слабый поток АТС. Все значения на рис. 7 и 8 усреднены за 10 мин.

6 Обсуждение результатов

В работе были изложены два алгоритма агрегации данных: на автомагистрали и на въездах и съездах. Для обоих алгоритмов были поставлены эксперименты на реальных данных с целью проверки их работоспособности. Таким образом, получены следующие результаты:

(a)

(б)

(в)

Рис. 6 Плотность АТС для результатов обучения модели (6) (a), (7) (б) и (8) (в)

Таблица 1 Сравнение моделей при всех значениях плотности ρ_{det}

Модель	Среднеквадратическая ошибка	Корреляция
(5)	0,03	0,787
(6)	0,031	0,781
(7)	0,0326	0,765
(8)	0,0314	0,78

Таблица 2 Сравнение моделей при плотности $\rho_{\text{det}} > 0,2$

Модель	Среднеквадратическая ошибка
(5)	0,058
(6)	0,119
(7)	0,065
(8)	0,093

Рис. 7 Пример восстановления числа въехавших АТС. Синяя линия — число проехавших под датчиком на въезде АТС, зеленые точки — сумма данных с датчика и GPS-треков в моменты времени из множества I_{in} . Красная линия — восстановленные значения суммарного числа въехавших АТС N_{estin} **Рис. 8** Пример восстановления числа въехавших АТС. Синяя линия — число проехавших под датчиком на въезде АТС, зеленые точки — сумма данных с датчика и GPS-треков в моменты времени из множества I_{in} . Красная линия — восстановленные значения суммарного числа въехавших АТС N_{estin}

1. Предложен алгоритм восстановления характеристик потока АТС по данным GPS-треков с использованием данных дорожных датчиков на автомагистрали.
2. Проведено сравнение различных моделей оценки числа АТС на автомагистрали с использованием реальных данных и выбрана лучшая.
3. Предложен алгоритм восстановления характеристик потока АТС по данным GPS-треков и дорожных датчиков с учетом возможных значительных временных пробелов в них на въездах и съездах с автомагистрали.
4. Проверена работоспособность предложенного алгоритма для восстановления суммарного потока на въездах и съездах на реальных данных.

Особенностью поставленной задачи оценки числа АТС на автомагистрали является большой объем данных с GPS-треков, что позволяет получить приемлемое качество на обучающей и контрольной выборке как для исходной задачи, так и для задачи в реальном времени.

В то же время из-за недостаточного количества данных на въездах и съездах нельзя решить задачу, аналогичную задаче на автомагистрали. Однако, используя уравнение баланса, можно оценить N_{in} и N_{out} на въездах и съездах. В работе предложен алгоритм для получения этих оценок, который дает интерпретируемые результаты.

7 Благодарности

Авторы благодарят компанию Яндекс (ООО «Яндекс») за предоставленные трековые данные, Центр организации дорожного движения за данные датчиков движения. Авторы также чрезвычайно признательны Я. А. Холодову и А. Е. Алексеенко за консультации и обсуждение результатов.

Литература

- [1] Алексеенко А.Е., Холодов Я.А., Холодов А.С., Горева А.И., Васильев М.О., Чехович Ю.В., Мишин В.Д., Старожилец В.М. Разработка, калибровка и верификация модели движения трафика в городских условиях. Часть I // Компьютерные исследования и моделирование, 2015. Т. 7. №6. Р. 1185–1203.
- [2] Бродский Г.С., Кашкин М.Ю., Айвазов А.Р., Рыкунов Р.Р. Работа детекторов транспорта на Московской дорожно-уличной сети URL: http://www.againc.net/media/3649/dt_article001.pdf.
- [3] Lighthill M.J., Whitham G.B. On kinematic waves. II. A theory of traffic flow on long crowded roads // P. Roy. Soc. Lond. A Mat., 1955. P. 281–345. doi: 10.1098/rspa.1955.0089.
- [4] Richards P.I., Shock waves on the highway // Oper. Res., 1956. Vol. 4. No.1. P. 42–51. doi: 10.1287/opre.4.1.42.
- [5] Whitham J.B. Linear and nonlinear waves. — USA: Wiley, 1974. 656 p.
- [6] Daganzo C.F. Requiem for second-order fluid approximations of traffic flow // Transport. Res. B Meth., 1995. Vol. 29. No. 4. P. 277–286. doi: 10.1016/0191-2615(95)00007-Z.
- [7] Payne H.J. Models of freeway traffic and control // Math. Models Public Syst., 1998. No. 4. P. 51–61.
- [8] Papageorgiou M. Some remarks on macroscopic traffic flow modelling // Transport. Res. A Pol., 1998. Vol. 32. No. 5. P. 323–329. doi: 10.1016/S0965-8564(97)00048-7.
- [9] Aw A., Rascle M. Resurrection of “second order” models of traffic flow // SIAM J. Appl. Math., 2000. Vol. 60. No. 3. P. 916–938. doi: 10.1137/S0036139997332099.
- [10] Zhang M. A non-equilibrium traffic model devoid of gas-like behavior // Transport. Res. B Meth., 2002. Vol. 36. No. 3. P. 275–290. doi: 10.1016/S0191-2615(00)00050-3.

- [11] *Zhang M.* Anisotropic property revisited —does it hold in multi-lane traffic? // *Transport. Res. B Meth.*, 2003. Vol. 37. No. 6. P. 561–577. doi: 10.1016/S0191-2615(02)00030-9.
- [12] *Siebel F., Mauser W.* On the fundamental diagram of traffic flow // *SIAM J. Appl. Math.*, 2006. Vol. 66. No. 4. P. 1150–1162. doi: 10.1137/050627113.
- [13] *Siebel F., Mauser W.* Synchronized flow and wide moving jams from balanced vehicular traffic // *Phys. Rev. E*, 2006. Vol. 73. No. 6. P. 066108. doi: 10.1103/PhysRevE.73.066108.
- [14] *Воронцов К.В., Чехович Ю.В.* Интеллектуальный анализ данных в задачах моделирования транспортных потоков // *Введение в математическое моделирование транспортных потоков* / Под ред. А. В. Гасникова. — М.: Изд-во МЦНМО, 2013. С. 225–248.
- [15] *Hall D.L., McMullen S.* Mathematical techniques in multisensor data fusion. — L.: Artech House, 2004. 453 p.
- [16] *Воронин А., Михеев Ю.* Синергетические методы комплексирования в задачах принятия решений // 12th Conference (International) “Knowledge–Dialogue–Solution” Proceedings. — Sofia: FOI-COMMERCE, 2006. P. 180–186.
- [17] *Khaleghi B., Khamis A., Karray F., Razavi S.* Multisensor data fusion: A review of the state-of-the-art // *Inform. Fusion*, 2013. Vol. 14. No. 6. P. 28–44. doi: 10.1016/j.inffus.2011.08.001.
- [18] *Goodman I., Mahler R., Nguyen H.* Mathematics of data fusion. — New York, NY, USA: Springer Science & Business Media, 2013. 507 p.
- [19] *Зенченко В.А., Ременцов А.Н., Павлов А.В., Сотсков А.В.* Оценка параметров окружающей среды и основных транспортных потоков, определяющих ситуацию на улично-дорожной сети // *Современные наукоемкие технологии*, 2012. № 2. С. 52–59.
- [20] *Jerry L., Ray D.* Violin plots: A box plot-density trace synergism // *Am. Stat.*, 1998. Vol. 52. No. 2. P. 181–184. doi: 10.1080/00031305.1998.10480559.
- [21] *Kerner B.* The physics of traffic. — Berlin: Springer, 2004. 681 p. doi: 10.1007/978-3-540-40986-1.

Поступила в редакцию 04.08.2016

Aggregation of data from different sources in traffic flow tasks*

V. M. Starozhilets^{1,2} and Yu. V. Chekhovich^{1,2}

starvsevol@gmail.com; chehovich@forecsys.ru

¹Federal Research Center “Computer Science and Control” of RAS

44/2 Vavilova Str., Moscow, Russia

²Moscow Institute of Physics and Technology

9 Institutskiy Per., Dolgoprudny, Moscow Region, Russia

Data aggregation problem, where data are taken from GPS-tracks and traffic detectors, has been studied. Aggregated data are used to state and solve finite differences equation corresponding to the chosen traffic flow mathematical model. The problem is divided into two ones: the first one is about highway data and the second one is about entrances and exits data. To estimate speed and number of cars, a linear model that uses highway data taken from GPS-tracks and traffic detectors has been proposed. The quality criteria are mean squared error and correlation coefficient. Note that the built model can be used on highway data, which do not have data from traffic detectors, but have only data from GPS-tracks. For entrances and

*This research is funded by the Russian Foundation for Basic Research, grant 14-07-00685

exits data, a method to recover summary total flow was developed. This method is based on the preservation of cars in transport network. Computational experiment for both problems is provided on real data and performance of the proposed approaches is demonstrated. Data from GPS-tracks were provided by Yandex.Traffic and data from traffic detectors were provided by Moscow traffic management center. Moscow Ring Road was used as a highway.

Keywords: *data recovery; Moscow Ring Road; GPS-tracks; road detectors*

DOI: 10.21469/22233792.2.3.01

References

- [1] Alekseenko, A. E., Ya. A. Kholodov, A. S. Kholodov, A. I. Goreva, M. O. Vasil'ev, Yu. V. Chekhovich, V. D. Mishin, and V. M. Starozhilets. 2015. Razrabotka, kalibrovka verifikatsiya modeli dvizheniya trafika v gorodskikh usloviyakh. Chast' I [Development, calibration and verification of the model of traffic in urban environments. Part I]. *Komp'yuternye issledovaniya i modelirovanie* [Computer Investigations and Simulation] 7(6):1185–1203.
- [2] Brodskiy, G. S., M. Yu. Kashkin, A. R. Ayvazov, and R. R. Rykunov. Rabota detektorov transporta na Moskovskoy dorozhno-ulichnoy seti [Traffic detectors working on the Moscow road network]. Available at: http://www.againc.net/media/3649/dt_article001.pdf (accessed November 3, 2016).
- [3] Lighthill, M. J., and G. B. Whitham. 1955. On kinematic waves. II. A theory of traffic flow on long crowded roads // *P. Roy. Soc. Lond. A Mat.* 281–345. doi: 10.1098/rspa.1955.0089.
- [4] Richards, P. I. 1956. Shock waves on the highway. *Oper. Res.* 4(1):42–51. doi: 10.1287/opre.4.1.42.
- [5] Whitham, J. B. 1974. *Linear and nonlinear waves*. USA: Wiley. 656 p.
- [6] Daganzo, C. F. 1955. Requiem for second-order fluid approximations of traffic flow. *Transport. Res. B Meth.* 29(4):277–286. doi: 10.1016/0191-2615(95)00007-Z.
- [7] Payne, H. J. 1998. Models of freeway traffic and control. *Math. Models Public Syst.* 4:51–61.
- [8] Papageorgiou, M. 1988. Some remarks on macroscopic traffic flow modelling. *Transport. Res. A Pol.* 32(5):323–329. doi: 10.1016/S0965-8564(97)00048-7.
- [9] Aw, A., and M. Rascle. 2000. Resurrection of “second order” models of traffic flow. *SIAM J. Appl. Math.* 60(3):916–938. doi: 10.1137/S0036139997332099.
- [10] Zhang, M.. 2002. A non-equilibrium traffic model devoid of gas-like behavior. *Transport. Res. B Meth.* 36(3):275–290. doi: 10.1016/S0191-2615(00)00050-3.
- [11] Zhang, M.. 2003. Anisotropic property revisited — does it hold in multi-lane traffic? // *Transport. Res. B Meth.* 37(6):561–577. doi: 10.1016/S0191-2615(02)00030-9.
- [12] Siebel, F., and W. Mauser. 2006. On the fundamental diagram of traffic flow. *SIAM J. Appl. Math.* 66(4):1150–1162. doi: 10.1137/050627113.
- [13] Siebel, F., and W. Mauser. 2006. Synchronized flow and wide moving jams from balanced vehicular traffic. *Phys. Rev. E* 73(6):066108. doi: 10.1103/PhysRevE.73.066108.
- [14] Vorontsov, K. V., and Yu. V. Chekhovich. 2015. Intellectual'nyy analiz dannykh v zadachakh modelirovaniya transportnykh potokov [Data mining for traffic flows modeling]. *Vvedenie v matematicheskoe modelirovanie transportnykh potokov* [Introduction to the mathematical modeling of traffic flows]. Ed. A. B. Gasnikov. Moscow: MTsNMO. 225–248.
- [15] Hall, D. L., and S. McMullen. 2004. *Mathematical techniques in multisensor data fusion*. L.: Artech House. 453 p.

- [16] Voronin, A., and Yu. Mikheev. 2006. Sinergeticheskie metody kompleksirovaniya v zadachakh prinyatiya resheniy [Synergetic methods of aggregation in decision problems]. *12th Conference (International) "Knowledge-Dialogue-Solution" Proceedings*. Sofia: FOI-COMMERCE. P. 180–186.
- [17] Khaleghi, B., A. Khamis, F. Kararay, and S. Razavi. 2013. Multisensor data fusion: A review of the state-of-the-art. *Inform. Fusion* 14(6):28–44. doi: 10.1016/j.inffus.2011.08.001.
- [18] Goodman, I., R. Mahler, and H. Nguyen. 2013. *Mathematics of data fusion*. New York, NY: Springer Science & Business Media. 507 p.
- [19] Zenchenko, V. A., A. N. Rementsov, A. V. Pavlov, and A. V. Sotskov. 2012. Otsenka parametrov okruzhayushchey sredy i osnovnykh transportnykh potokov, opredelyayushchikh situatsiyu na ulichno-dorozhnoy seti [Evaluation of the environmental parameters and the main traffic flows that determine the situation on the road network]. *Sovremennye Naukoemkie Tekhnologii* [Modern High Technologies] 2:52–59.
- [20] Jerry, L., and D. Ray. 1998. Violin plots: A box plot-density trace synergism. *Am. Stat.* 52(2):181–184. doi: 10.1080/00031305.1998.10480559.
- [21] Kerner, B. 2004. *The physics of traffic*. Berlin: Springer. 681 p. doi: 10.1007/978-3-540-40986-1.

Received August 4, 2016