Московский физико-технический институт
(Государственный университет)

Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

ДИПЛОМНАЯ РАБОТА СТУДЕНТА 974 ГРУППЫ

«Прогнозирование времени прибытия автомобильных рейсов, совершающих междугородние перевозки»

Выполнил:
студент 4 курса 974 группы
Шульга Александр Вадимович

Научный руководитель:
к. ф-м. н., н. с. ВЦ РАН
Чехович Юрий Викторович

Москва, 2013

Содержание

1 Введение **3**

2 Постановка задачи, исходные гипотезы, входные данные **9**

2.1 Основные обозначения и исходные гипотезы 9

2.2 Постановка задачи 10

2.3 Задача минимизации 11

2.4 Структура данных 11

2.5 Предобработка данных 12

3 Описание предложенных алгоритмов **12**

4 Вычислительный эксперимент **15**

4.1 Построение историй проездов H_i 15

4.2 Алгоритм сложения средних времен проездов отрезков BaseAlg 17

4.3 Алгоритм сложения гистограмм HistAlg 18

4.4 Алгоритм построения эмпирической функции распределения методом сэмплирования DistrAlg 19

5 Заключение **23**

Аннотация

Рассмотрена задача прогнозирования времени прибытия автомобилей, совершающих междугородние перевозки. Были предложены алгоритмы, основанные на истории проездов автомобилей, которые решают данную задачу на выделенном участке дороги Тосно-Химки. Был сделан сравнительный анализ и предложены критерии качества каждого из алгоритмов. Итоговая версия прогноза основана на построении эмпирической функции распределения по временам проездов по всей дороге, используя метод сэмплирования для функций распределения отдельных участков дороги. Исследованы устойчивость итогового алгоритма к многократным запускам и к прореживанию исходных данных. Так же была исследована его обобщающая способность. Все алгоритмы тестировались на реальных данных.

Ключевые слова: *метод сэмплирования, travel-time prediction, функция распределения случайной величины.*

1 Введение

Актуальность темы. На сегодняшний день задача прогнозирования времени прибытия автомобилей является весьма существенной для различных прикладных областей. Давно известно, что пробки на дорогах в больших городах, а так же на основных магистралях и дорогах, переросли в проблему едва ли не первой степени важности. Предсказание времени прибытия автомобилей может быть полезным для многих компаний, связанных с транспортом. Дальнейшие исследования в этой сфере могут значительно упростить и улучшить как транспортные перевозки, так и пассажирские.

Цель работы. Построить несколько вариантов алгоритмов, основанных на исторических данных проездов, которые будут решать поставленную задачу прогнозирования времени прибытия автомобилей. Построить критерии для оценки качества прогнозирования алгоритмов.

Методы исследований. При построении алгоритмов использовались методы предобработки данных — медианная фильтрация, фильтрация выбросов и неинформативных треков, методы восстановления плотности распределения, методы построения эмпирических функций распределения случайных величин и суммы случайных величин. При оценке качества алгоритмов использовались методы сравнения эмпирических функций распределения и методы оценивания абсолютных ошибок алгоритмов. Для программной реализации разработанных алгоритмов использовались среды MATLAB, MySQL.

Научная новизна.

- Предложена модель для метода сэмплирования, которая учитывает корреляции времен проездов соседних участков дороги.
- Разработан алгоритм на основе модели, который выдает не единичное значение прогноза времени, а распределение вероятности по времени прибытия.
- Качество прогноза данного алгоритма с накоплением истории проездов будет улучшаться.

Практическая ценность. Разработаны части программного модуля, который:

- Делает предобработку входных данных;
- Разбивает заданный маршрут на отрезки;
- Привязывает к каждому отрезку историю проездов по нему;
- Строит алгоритм на основе историй проездов по каждому из отрезков;
- Визуализирует результаты.

Положения, выносимые на защиту:

- Алгоритм прогнозирования времени прибытия автомобилей, основанный на построении функции распределения суммы случайных величин — времен проездов по отрезкам дороги, с помощью метода сэмплирования.

Апробация. Результаты квалификационной работы бакалавра были использованы для решения задачи прогнозирования времени прибытия автомобилей для компании Cargo — РНТ.

Обзор литературы. [1]. Short Term Traffic Prediction Models
Обзорная статья по трем основным типам моделей для предсказания трафика — наивные, параметрические и непараметрические. Приводятся их основные разновидности, вкратце рассказывается про ключевые недостатки и преимущества. В конце статьи дана сравнительная таблица, а так же огромный список литературы.

[2].Inter - Urban Short Term Traffic Congestion Prediction
В статье представлено достаточно много теоретического материала по тематике транспорта. Дан обзор различным алгоритмам, так же есть история DTM [3]. Но практической ценности статья не представляет, так как нет никаких исследований и проектов.

[4].Прогнозирование загруженности автомобильных дорог

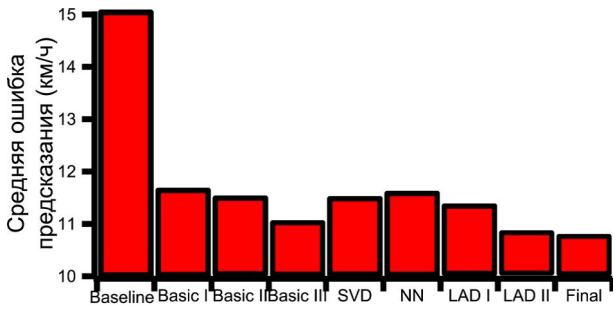

Рис. 1: Сравнительная статистика алгоритмов по метрике Яндекса

Рассматриваются различные подходы к предсказанию трафика, и делается сравнительная статистика работы различных методов. Основные рассмотренные методы — средневзвешенное, сингулярное разложение, нейронные сети, метод ближайших соседей, а так же их модификации (кластеризация, глобальные эффекты и комбинирование различных методов).Вкратце изложены основные идеи и формулы алгоритмов.

Результаты работы представлены на Рис. 1, где:

Baseline — это «простая скептическая оценка»: средняя скорость для участка по всем дням месяца в одно и тоже время. Если данных не было, то скорость полагали равной 0;

Basic I, II, III — базовые модели I, II и III соответственно;

SVD — алгоритм, на основе метода сингулярного разложения;

NN — алгоритм, на основе нейронных сетей;

LAD I — комбинирование результатов вышеупомянутых алгоритмов, за исключением базовой модели III;

LAD II — комбинирование результатов базовой модели III и LAD I;

Final — финальный результат, построенный на основе базовой модели III с использованием кластеризации.

Используя данные Яндекса, делается вывод, что простые алгоритмы, основанные на вычислении средних скоростей, работают лучше алгоритмов, построенных на основе сложных моделей. Минусом статьи является отсутствие сравнения работы алгоритмов на других данных, а так же не проводится сравнение с алгоритмами, у которых больше параметров. Данная статья может быть полезна при использовании в прогнозировании фактора загруженности дорог.

[5].Real-time travel path prediction using GPS-enabled mobile Phones

Работа посвящена предсказанию маршрута в реальном времени с помощью GPS на телефоне. Предложен алгоритм (с графической схемой), который использует исторические данные и положение в реальном времени. Алгоритм состоит из двух частей. Первая - строит путь по данным от GPS. Вторая часть алгоритма делает прогноз времени, основываясь на сходстве шаблонов.

Результаты : Алгоритм тестировался на сети Sprint - Nextel CDMA и было показано, что он выдает правильные результаты. Таким образом, данная концепция алгоритма работает. Минусы алгоритма в том, что он нуждается в доработке, потому что в нем задействовано мало параметров. Никаких конкретных численных данных нет. Так же нет готовых продуктов.

[6].Travel-Time Prediction using Gaussian Process Regression: A Trajectory-Based Approach

Статья про алгоритм, использующий регрессию для гауссовского процесса (Gaussian Process Regression). Новизной является способность предсказывать время для любого пути (в том числе и не использовавшегося ранее), если схожесть между участками путей определена с помощью ядра (kernel function [7]). При этом не нужно использовать большие массивы данных. Для ускорения вычислений GPR используется факторизация Холецкого [8].

Результаты - Эксперимент проводился на реальных данных центра г.Киото, Япония и дал весьма точные предсказания времени. В минусы можно записать большое время вычислений. Так же было рассмотрено мало видов ядер.

[9].Travel Time Prediction System (TIPS)

Статья про оценивание эффективности внедренной системе прогнозирования TIPS, которая использует свои детекторы. В статье представлено достаточно много

таблиц и графиков, но теоретических расчетов не много. В результате было установлено, что при прогнозировании с использованием временных рядов с лагом в 30 секунд, предсказанные значения должны отображаться на Changeable Message Sign (CMS) для водителей с 4х минутными интервалами, и информация на CMS не должна меняться в течении 3х минут.

[10].Traffic Modelling and Prediction using ARIMA/GARCH Model

Статья о комбинации двух моделей — ARIMA и GARCH [11]. Весьма подробно описана сама модель и методы оценивания ее параметров. В сравнении ставилась модель FARIMA [12]. Тестировались они на реальных данных LBL-TCP-3.

Основные результаты - нелинейные временные ряды могут делать прогнозы лучше, чем классические линейные, даже если линейные временные ряды могут вести себя автомодельно (selfsimilarity). Новизна работы в том, что модель позволяет работать с нестационарными временными рядами. В то же время есть и проблемы - неустойчивость и сложность вычислений.

[13].Evaluation of DynaMIT - A Prototype Traffic Estimation and Prediction System

Еще один готовый продукт — DynaMIT . Статья о внедрении системы DynaMIT в Hampton Roads network. Параметрами модели являются матрицы корреспонденций. Статья большая, с многочисленными выкладками и графиками.

Основные результаты — DynaMIT показала хорошую производительность с RMSN ошибками в пределах $0.15 - 0.25$ для оценивания traffic sensor, когда эти предсказанные traffic sensor counts были в пределах $0.25 - 0.4$. Проблемы DynaMIT: подсчет трафика — не очень чувствительный параметр в оценивании эффективности и прогнозировании системы. К тому же дополнительные параметры могут быть неоптимальными для всех сегментов.

[14].An Adaptive Travel Time Prediction Model Based On Pattern Matching

В данной статье предложена модель предсказания трафика, основанная на сходстве шаблонов. «Похожесть» находят с помощью генетического алгоритма (который весьма подробно описан). Данные проверялись на inbound section of route no. 3 of the Tokyo Metropolitan Expressway, i.e. from Yoga to Tanimachi. Результаты — адаптивные параметры улучшают оценки модели и соответствуют интуиции (лучше, чем stepwise параметры). Средние абсолютные ошибки и средние абсолютные процентные ошибки и выше находятся в пределах — $\pm 5\%$, $\pm 10\%$ и ± 5 минут. Большим плюсом является то, что адаптивные параметры увеличивают эффективность модели, и она может быть внедрена для любой дороги с любыми условиями без модификаций.

[15].Short-Term Traffic Forecasting Using The Local Linear Regression Model

Небольшая статья про краткосрочное прогнозирование трафика с использованием модели локальной линейной регрессии. Достаточно подробно описан алгоритм. Тестировался он на Houston US-290 Northwest freeway eastbound .

Результаты — данный алгоритм более эффективен, чем kNN, локальные константные методы и ядерное сглаживание. Так же, в отличие от этих методов, LLR использует как исторические, так и текущие данные. Проблемы данного метода — сильное различие при одно- и многошаговом предсказании.

[16].Как работают Яндекс.Пробки

Небольшая статья собственно о том, как работают Яндекс.Пробки. Описан только поэтапный план, никаких теоретических выкладок нет. Данные собираются с различных источников, отсеиваются ненужные, и потом определяется загруженность транспортной сети. Далее составляется карта пробок города. А время маршрута определяется в сравнении с эталонным временем проезда по данному участку. Плюсы данного проекта — простота реализации. Минусы — большая неточность прогноза, и результаты предоставлены только в пределах одного города.

[17].Расчет оптимального маршрута 2.0

Короткая обзорная статья о том, как при расчете оптимального маршрута используется Google.Maps API. Никакого теоретического материала или даже алгоритма нет. Так как при определении времени маршрута нужен сам трек, то все недостатки этого сервиса у Google наследуются от недостатков Google.Maps API. Плюсами проекта является то, что он хорошо прокладывает маршрут в рамках одного города и около $8-10$ промежуточных пунктов. Минусы — мало промежуточных пунктов. Так же проект не умеет строить маршруты между городами.

2 Постановка задачи, исходные гипотезы, входные данные

2.1 Основные обозначения и исходные гипотезы

Задано:

- Точка — положение машины в определенное время;
- Трек $\mathcal{T}$ — последовательность точек данного автомобиля;
- Маршрут $\mathcal{M}$ — трек от стартовой точки A до конечной B;
- Отрезок i — ребро графа дорог $\mathcal{G}$;
- История проездов H — таблица со всеми данными;

Введено:

- H_i — таблица времен проездов всех автомобилей по данному отрезку дороги;
- $\bar{\mathbf{x}} = \bar{\mathbf{x}}(t) = (x, \bar{\theta}(t))$ — объект, где x — описание автомобиля, $\bar{\theta}(t)$ —параметры автомобиля в каждый момент времени.
- T_i — время въезда на i -й отрезок;
- t_i — время проезда i -го отрезка;
- T — общее время в пути.

Таблица треков H задает признаковое описание объектов, т.е. объект — автомобиль, совершающий маршрут из точки A в заданное известное время в точку B.

Описание объекта — последовательность точек трека данного автомобиля.
Пусть $\bar{\mathbf{x}} = \bar{\mathbf{x}}(t)$ — объект. Его можно представить в виде $\bar{\mathbf{x}} = (x, T_0, \bar{\theta}(t))$, где x — описание автомобиля, $\bar{\theta}(t)$ —параметры автомобиля в каждый момент времени (его местоположение).

Основное предположение — истории проездов H достаточно для построения прогноза.

Требования, предъявляемые к модели.

- Адекватное соответствие реальному времени проезда по выбранному маршруту:
  $|t_{real} - t_{forecast}| \leq \varepsilon$.
- Устойчивость прогноза к разреженности начальных данных, а так же к многократному запуску алгоритма;
- Отсутствие переобучения (желательно). Алгоритм должен при тех же параметрах работать для других участков дороги.

2.2 Постановка задачи

Дано:

- множество отрезков заданного маршрута $\mathcal{M} : \{i\}, i = \overline{1, n}$;
- история проездов каждого из отрезка H_i;
- Объект $\bar{\mathbf{x}}$;
- Контрольная выборка $X^k = \{\bar{\mathbf{x}}_\mathbf{1}, \ldots, \bar{\mathbf{x}}_\mathbf{k}\}$ с ответами в виде времен проездов всего маршрута $Y^k = \{T_1, \ldots, T_k\}$.

При вычислении времени въезда на отрезок предполагается преобразование типов данных.

Требуется построить функцию f:

$$t_i = f(\bar{\mathbf{x}}_\mathbf{j}(T_i), T_i, H_i, \Theta),$$

где Θ — параметры функции f.

Тогда прогноз для всего маршрута будет:

$$T = F(\bar{\mathbf{x}}_\mathbf{j}, T_0, H_1 \dots H_n, \Theta),$$

где

$$F(\bar{\mathbf{x}}_\mathbf{j}, T_0, H_1 \dots H_n, \Theta) = \sum_{i=1}^{i=n} f(\bar{\mathbf{x}}_\mathbf{j}(T_i), T_i, H_i, \Theta),$$

$$T_{i+1} = T_i + f(\bar{\mathbf{x}}_\mathbf{j}(T_i), T_i, H_i, \Theta).$$

Вводится функция потерь $\mathcal{L}(T_i, F(\bar{\mathbf{x}}_\mathbf{j}, T_0, H_1 \dots H_n, \Theta))$ и функционал качества алгоритма

$$Q(f, X^k) = \sum_{j=1}^{j=k} \mathcal{L}(T_i, F(\bar{\mathbf{x}}_\mathbf{j}, T_0, H_1 \dots H_n, \Theta)).$$

2.3 Задача минимизации

$$f = \arg\min_{\Theta} Q(F(\bar{\mathbf{x}}_\mathbf{j}, T_0, H_1 \dots H_n, \Theta), X^k).$$

Таким образом исходная проблема сводится к двум следующим подзадачам:

1. Выбор алгоритма f — задать модель на данных H и H_i
2. Оптимизировать параметры функции Θ.

В данной работе будет предложено несколько вариантов решения этих двух подзадач.

2.4 Структура данных

История проездов автомобилей представлена в виде таблицы треков H, полученных по данным GPS:

- **vehicle_id** — id автомобиля;
- **id** — id точки;
- **date** — время с точностью до 2х секунд;
- **latitude** — широта;
- **longitude** — долгота;

История проездов по каждому из отрезков представлена в виде таблицы H_i:

- **vehicle_id** — id автомобиля;
- **id_tr** — id маршрута данного автомобиля;
- **date** — время въезда на данный отрезок точностью до 2х секунд;
- **t** — время проезда по данному отрезку;

Граф дорог был выделен из карт OSM [18] и представляет собой набор точек с координатами а так же участки дорог в виде последовательностей точек. Направление участка дороги задается параметрами в тэгах а так же последовательностью точек. Участки дорог не всегда являются ребрами графа, поэтому отдельной задачей является выделение ребер из $\mathcal{G}$.

2.5 Предобработка данных

История проездов не содержит разбиения на маршруты, поэтому пришлось разбивать на маршруты по критерию задержки сигнала. Если следующая точка в треке в H по времени была позже более чем на 3 часа — это начало нового маршрута.
Так же в H присутствовали как повторяющиеся записи, так и выбросы, связанные с неточностью GPS, поэтому проводилась фильтрация данных.
Выбранный участок дороги от Тосно до Химок разбивался на отрезки i. Для каждого отрезка выделялись все автомобили, которые проезжали эти отрезки, и вычислялись времена проездов. Таким образом были сформированы истории проездов H_i.
При привязке истории проездов H_i к i -му отрезку делалась фильтрация остановок и быстрых времен проездов (связанные с ошибками в данных) с помощью медианной фильтрации с порогами в 10 %. Так же для каждой точки в таблице H вычислялось расстояние до дороги и фильтровались точки, которые лежат дальше, чем значение порога h.

3 Описание предложенных алгоритмов

- **Алгоритм сложения средних времен проездов отрезков(далее BaseAlg)**: Для каждого отрезка $i = \overline{1,n}$ вычисляется среднее значение времени проезда t_i по подмножеству $H_{\bar{\theta},i} \subseteq H_i$ в зависимости от входных параметров $\bar{\theta}$ объекта $\bar{\mathbf{x}}$. Итоговый прогноз — $T(\bar{\mathbf{x}}) = \sum_{i=1}^{i=n} t_i$.

Вход: входные параметры $\Theta_0 = \bar{\theta}(T_0)$, число отрезков n

Выход: прогноз времени T

1. **Для всех** $i = 1, \ldots, n$
2. Вычислить среднее время проезда для отрезка i — t_i по подвыборке $\tilde{H}_i = H_i(\Theta_i) = H_{\bar{\theta},i}$;
3. Пересчитать входные параметры для следующего отрезка $\Theta_{i+1} = \bar{\theta}(T_{i+1})$ и время въезда на следующий отрезок T_{i+1};
4. $T := T + t_i$;

- **Алгоритм сложения гистограмм (далее HistAlg)**:

  Задается параметр h — шаг гистограммы, одинаковый для всех отрезков. Для каждого отрезка $i = \overline{1,n}$ вычисляется нормированная гистограмма G_i распределения времен проездов в соответствии с величиной h по подмножеству $H_{\bar{\theta},i} \subseteq H_i$. G_i представляет собой таблицу с двумя столбцами t и p, где t — временной интервал ширины h (интервал можно отождествить, например, с его серединой), а p — доля значений времен проездов, попавших в интервал t;

  Предполагается, что отрезки независимы. Тогда формула для сложения гистограмм имеет вид:

  $G_{1+2} = \{(t_{min}, p(t_{min})) \ldots (t_{max}, p(t_{max}))\}$;

  $t_{min} = min(t_1) + min(t_2),\ t_1 \in G_1, t_2 \in G_2$;

  $t_{max} = max(t_1) + max(t_2),\ t_1 \in G_1, t_2 \in G_2$;

  $p_{1+2}(t) = \sum\limits_{t_1+t_2=t} p_1(t_1) \cdot p_2(t_2),\ \forall t \in G_{1+2}$;

  Вход: входные параметры $\Theta_0 = \bar{\theta}(T_0)$, число отрезков n, ширина окна h

  Выход: прогноз времени T

  1. **Для всех** $i = 1, \ldots, n$
  2. Построить нормированную гистограмму $G_i = \{t, p_i(t)\}$ времени проезда по подвыборке $\tilde{H}_i = H_i(\Theta_i) = H_{\bar{\theta},i}$;

3. **Выход из цикла**
4. Применить алгоритм сложения гистограмм соседних отрезков для всех отрезков с получением итоговой гистограммы G_T

5. Построить прогноз по G_T.

Итоговый прогноз представляет собой гистограмму — распределение вероятности по временным интервалам.

Так как параметр h может быть очень малым, итоговую гистограмму можно пересчитать для новых интервалов ширины $\hat{h} > h$ по формуле:

$p(\hat{t}) = p[\hat{t}, \hat{t} + \hat{h}] = \sum\limits_{t \in [\hat{t}, \hat{t}+\hat{h}]} p(t).$

- **Алгоритм построения эмпирической функции распределения методом сэмплирова**

Для каждого отрезка $i = \overline{1, n}$ вычисляется эмпирическая функция распределения $F_i(t)$, где t —время проезда, по подмножеству $H_{\bar{\theta},i} \subseteq H_i$. Затем вычисляются коэффициенты корреляции c_i для каждых соседних пар отрезков $[i-1, i]$.

Известно, что если $t \in F,\ r \in U(0,1) \Rightarrow F_t^{-1}(r) \in F$.

Суть метода сэмплирования заключается в том, что N раз для каждого отрезка выбираются числа r_i, вычисляются значения $t_i = F_i^{-1}(r_i)$, а затем суммируются $T_j = \sum_{i=1}^{i=n} t_i,\ j = \overline{1, N}$. По найденным T_j строится итоговая эмпирическая функция распределения F_T.

Задача состоит в выборе модели подбора случайных величин $r_i = g(i, r_{i-1}, c_i)$.

Вводится порог p_c, для которого если $c_i < p_c,\ g(i, r_{i-1}, c_i) = \text{rand}(0,1)$.

Если $c_i \geq p_c$, то функция g вводится по следующим соображениям:

Так как коэффициент корреляции для соседних двух отрезков больше порога, значит должны коррелировать и времена проездов $t_i,\ t_{i-1}$, и, следовательно, r_i и r_{i-1}. Значение r_i должны выбираться случайным образом из интервала, который содержит $r_{i-1}:\ r_i \in [r_{i-1} - a, r_{i-1} + b]$, причем возможно, что $a \neq b$.

Величины a, b должны зависеть от коэффициента корреляции:

Соответственно, если $c_i = 1$ — линейная связь, то интервал «схлопывается» в точку $r_{i-1},\ a = b = 0$. Если же $c_i < p_c$, то интервал вырождается в $[0, 1]$.

Вход: входные параметры Θ_0, число отрезков n, порог коэффициента корреляции p_c, w_1, w_2

Выход: прогноз времени T

1. **Для всех** $i = 1, \ldots, n$
2. Для отрезка i построить эмпирическую функцию распределения $F_i(t)$ времени проезда по $\tilde{H}_i = H_i(\Theta_i)$;
3. Для отрезка i вычислить коэффициент корреляции c_i с отрезком $i-1$;
4. **Для всех** $j = 1, \ldots, N = 1000 \ldots 30000$
5. инициализировать $T_j := 0$
6. $r_1 = \text{rand}(0, 1),\ T_j := F_1^{-1}(r_1)$
7. **Для всех** $i = 2, \ldots, n$
8. $r_i = \text{rand}(r_{i-1} - a, r_{i-1} + b)$

   Если $c_i < p_c$, **Тогда** $a = 0,\ b = 1$

   Иначе $a = \varphi_1(r_{i-1}, c_i, \mathbf{w}_1),\ b = \varphi_2(r_{i-1}, c_i, \mathbf{w}_2)$.
9. $t_i = F_i^{-1}(r_i),\ T_j := T_j + t_i$
10. **Выход из цикла по i**
11. **Выход из цикла по j**
12. Построить эмпирическую функцию распределения F_T по T_j
13. Построить прогноз по F_T.

Прогноз по функции $F_T(t)$ можно построить либо вычислив функцию плотности $f_T(t)$, либо построив график распределения вероятности по интервалам времени, для которого вероятность попасть в интервал $[t, t+k]$:
$P[t, t+k] = F_T(t+k) - F_T(t)$.

4 Вычислительный эксперимент

4.1 Построение историй проездов H_i

Первым делом выделялась дорога $\mathcal{M}$ из Тосно в Химки из карт OSM. Нужно было упорядочить точки маршрута в порядке их следования. Для этого использовался поиск в глубину (предполагалось, что карты не содержат ошибок и тогда алгоритм

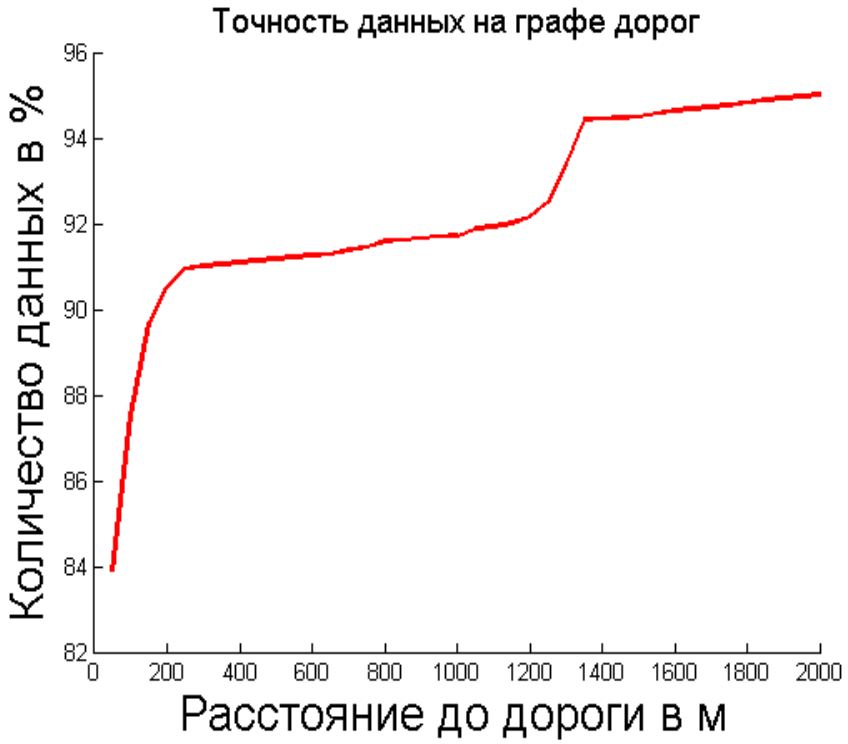

Рис. 2: Доля точек внутри полосы ширины h, в зависимости от h.

правильно находит последовательность). Всего точек в $\mathcal{M}$ около 2600.

Начальная таблица с треками H содержит около 80 млн записей. Данных о точных разбиениях на маршруты не было, поэтому пришлось разбивать «вручную» по задержке сигналов от GPS. Если следующая точка трека была на 3 часа позже текущей — это новый маршрут. Далее отбрасывались все данные, которые не попадали в условный прямоугольник, образованный самыми крайними точками маршрута $\mathcal{M}$. После фильтрации повторяющихся записей в итоге получилась первичнообработанная история проездов $\hat{H}$, которая содержит порядка 12 млн записей.

Для каждой точки A вычислялось расстояние до двух ближайших точек маршрута $\mathcal{M}$, а так же расстояние между этими точками. По формуле Герона находилось расстояние h от A до дороги.

График зависимости доли точек из истории, лежащих внутри полосы ширины h вокруг дороги от h приведен на Рис. 2.

Видно, что 90 % данных лежит в 200м полосе. Излом графика объясняется тем, что

при больших h в долю данных попадали точки и с других дорог, которые примыкают к $\mathcal{M}$. Остальные точки были отфильтрованы, как выбросы, связанные с неточностью GPS.

Далее, нужно было разбить исходную дорогу на отрезки. Выделялись все дороги, которые имели общие точки с $\mathcal{M}$. Тогда точки пересечения и являлись концами отрезков. По сути отрезок — это участок дороги, с которого нельзя съехать и на который нельзя въехать, кроме как на концах. Таким образом было получено 506 отрезков $i = \overline{1, 506}$.

Потом для каждой точки определялся отрезок, которому она принадлежит. Формально, точка принадлежит отрезку, если ее проекция на дорогу лежит внутри этого отрезка.

Вычисление историй проездов H_i производились следующим образом:
Если две соседние точки $(x_1, T_1), (x_2, T_2)$ трека (для данного маршрута данного автомобиля) лежат соответственно на i и k отрезке $(i < k)$, то вычислялась скорость $v = \frac{\text{dist}(x_1, x_2)}{T_2 - T_1}$, а затем время въезда на каждый из отрезков $\{T_j |\ j : j > i,\ j \leq k\}$ и время проезда каждого из отрезков $\{t_j |\ j : j > i,\ j < k\}$.
К этим таблицам проездов применялась медианная фильтрация с 10 % порогом и в итоге получались H_i.

4.2 Алгоритм сложения средних времен проездов отрезков BaseAlg

Задавался только один входной параметр — час недели $t = n_h + 24n_d$, где n_h — час дня, когда стартовала машина, а $n_d = \overline{0, 6}$ — номер дня.
На Рис. 3, 4 приведен график зависимости времени проезда от часа недели, а так же относительная ошибка в сравнении с контрольной выборкой. Ошибка вычисляется по формуле:
$\text{Err} = \frac{|f(t) - x(t)|}{x(t)}$.

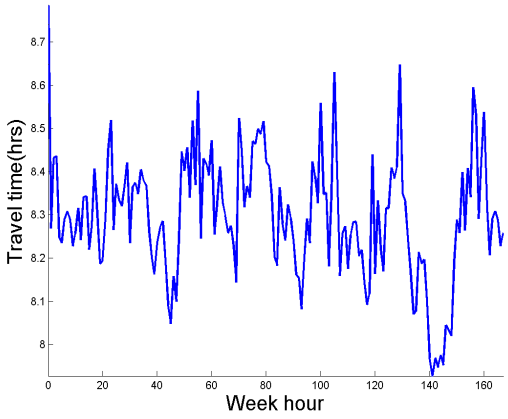

Рис. 3: Прогноз алгоритма **BaseAlg** в зависимости от часа недели.

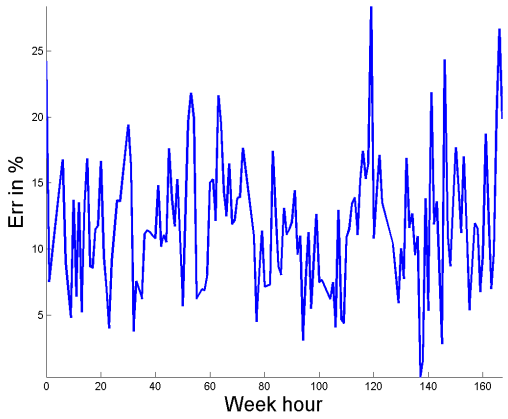

Рис. 4: Ошибка алгоритма **BaseAlg** в сравнении с контрольной выборкой в зависимости от часа недели.

Видно, что прогноз дает не достаточно качественные результаты и что ошибка данного алгоритма не превышает 30 %, но при этом значения сильно разбросаны.

4.3 Алгоритм сложения гистограмм HistAlg

В зависимости от входных параметров $\bar{\theta}$ (час или день недели, тип автомобиля, и т. д.), делается выборка истории $H_{\bar{\theta},i} \subseteq H_i$ для каждого из отрезков. По ней строится гистограмма доли выборки, попавшей во временной интервал t заданной ширины h. Предполагается, что соседние отрезки независимы.

Пример работы алгоритма для двух произвольных соседних отрезков представлен на Рис. 5, 6, 7, 8.

Контрольная выборка так же формировалась из автомобилей, проехавших весь маршрут. Но остановки на отрезках j заменялись на наиболее вероятные значения времен проездов в соответствии с гистограммой для j. На Рис. 9, 10 представлены итоговые гистограммы с величиной шага $h = 5$ с и $h = 10$ с в сравнении с контрольной выборкой. Пересчитывались данные гистограммы на интервалы шириной в 50 с.

В связи с тем, что вычисление итоговой гистограммы — весьма трудоемкая процедура, а сама итоговая гистограмма лишь приближает вид распределения вероятности по интервалам, алгоритм **HistAlg** так же является не эффективным: оптими-

зация параметра h будет занимать много времени, а решение будет неустойчивым. Идея же настройки параметров h_i для каждого из отрезков может привести как к еще большему увеличению трудоемкости, так и к возможному переобучению.

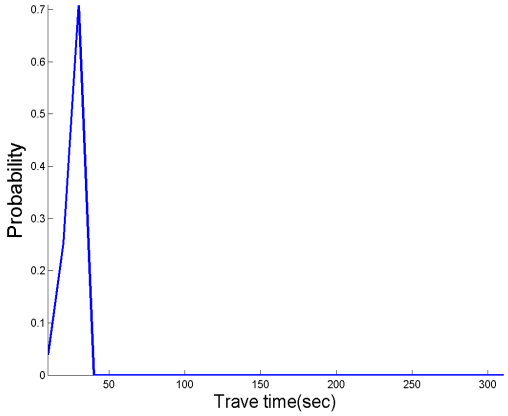

Рис. 5: Гистограмма первого отрезка.

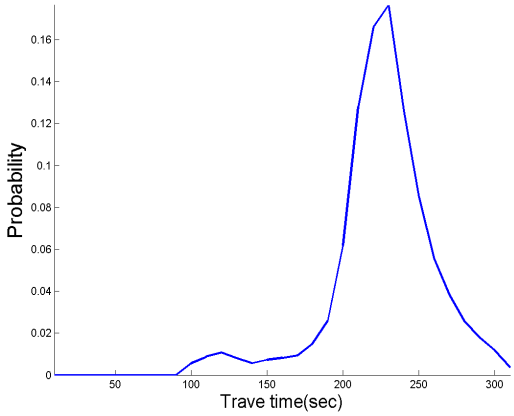

Рис. 6: Гистограмма второго отрезка.

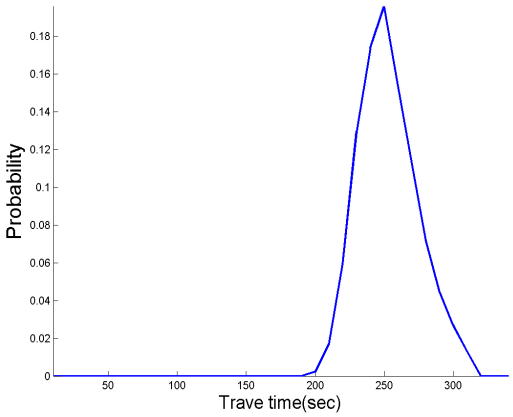

Рис. 7: Гистограмма объединения отрезков.

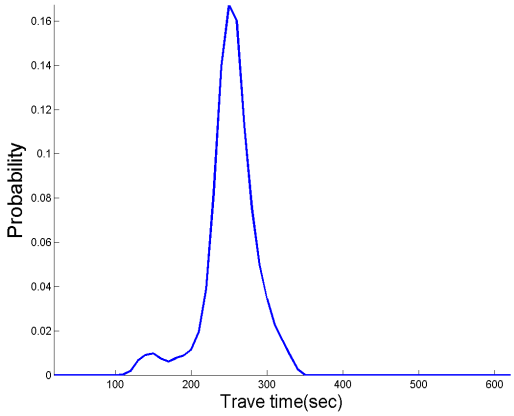

Рис. 8: Алгоритм сложения гистограмм.

4.4 Алгоритм построения эмпирической функции распределения методом сэмплирования DistrAlg

Чтобы еще раз показать, что Алгоритм сложения гистограмм не эффективен, предлагается сравнение функций распределений контрольной выборки, и прогноза, построенного в предположении о независимости соседних отрезков на Рис. 11.

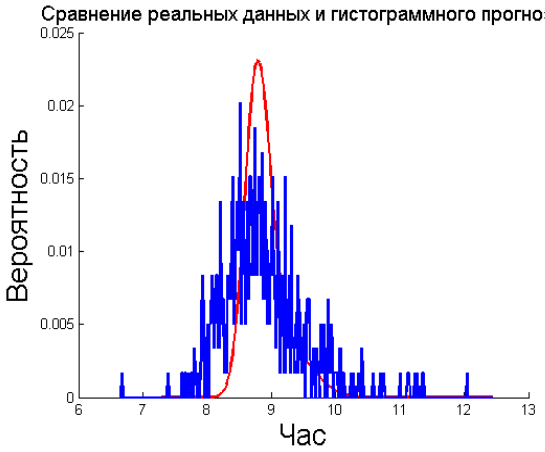

Рис. 9: **HistAlg** (красный) в сравнении с контрольной выборкой, $h = 5$с.

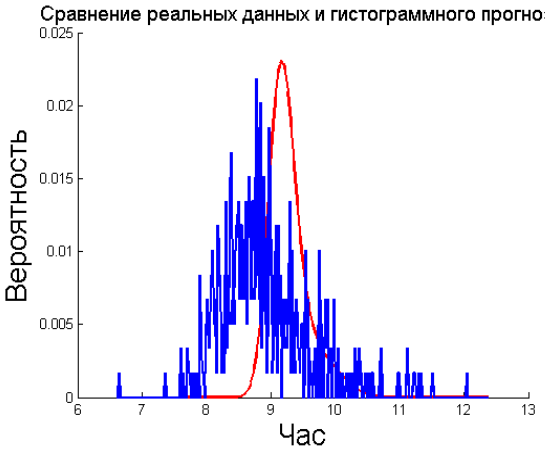

Рис. 10: **HistAlg** (красный) в сравнении с контрольной выборкой, $h = 10$с.

Зависимость вида функции распределения от порога p_c представлена на Рис. 12, 13, 14 и 15.

В результате подбора параметров эмпирическим методом, был установлен вид функции g:

If $c_i < p_c$, then $a = 0,\ b = 1$;

Else

If $r_{i-1} > 0.5,$ then

$a := \max\{r_{i-1} - (1 - r_{i-1})(1 - c_i)/0.815,\ 0\}$;

$b := \min\{(1 - r_{i-1}) + (1 - r_{i-1})(1 - c_i)/4,\ 1\}$;

Else

$a := \max\{r_{i-1} - (r_{i-1})(1 - c_i)/0.815,\ 0\}$;

$b := \min\{r_{i-1} + (r_{i-1})(1 - c_i)/4),\ 1\}$;

$\mathrm{g}(i, r_{i-1}, c_i) = \mathrm{rand}(r_{i-1} - a, r_{i-1} + b)$.

Сравнение графиков функций распределения итоговой модели и контрольной выборки представлено на Рис. 16.

Так же для анализа можно использовать методы статистики. В данной работе применялся критерий Колмогорова-Смирнова [19]. При текущих параметрах для $N = 3000$ (число сэмплирований) значение статистики Колмогорова-Смирнова (далее λ_α) составляет в среднем около 0.96.

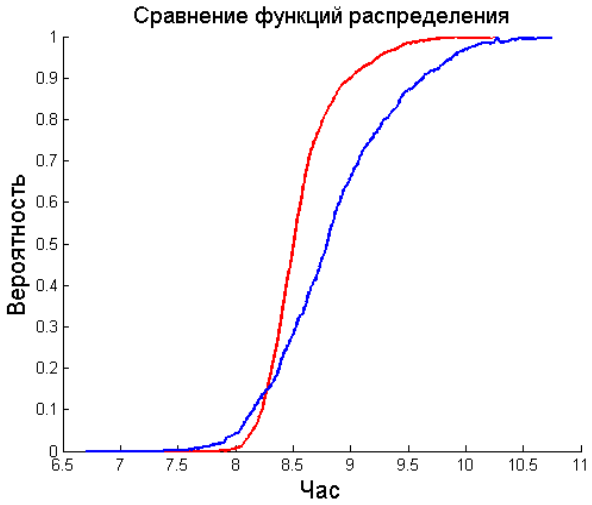

Рис. 11: Сравнение функции распределения прогноза (красным) с контрольной выборкой при гипотезе о независимости отрезков.

Это весьма хороший результат согласно таблице критических значений статистики Колмогорова-Смирнова:

Таблица 1: Критические значения статистики Колмогорова-Смирнова

α	0.20	0.10	0.05	0.02	0.01	0.001
λ_α	1.073	1.224	1.358	1.520	1.627	1.950

Полученную модель так же нужно было проверить на устойчивость. Первый способ — повторить построение N сэмплирований T раз и вычислить значения λ_α. Второй способ — истории проездов $H_{\bar{\theta},i} = \hat{H}_i$ разбить независимо на две произвольные подвыборки различной длины $\hat{H}_{i,less}$ и $\hat{H}_{i,bigg}$. По ним построить прогноз и вычислить значения статистики λ_α. Результаты экспериментов на устойчивость приведены в таблице 2 и 3.

Таблица 2: Устойчивость модели при многократном запуске сэмплирований

№ эксперимента	1	2	3	4	5	6	7	8	9	10
λ_α	0.805	0.985	0.805	1.202	0.598	0.989	0.859	1.206	1.081	1.079

Таблица 3: Устойчивость модели при переразбиении истории проездов H_i

less	1.857	1.954	0.867	1.064	1.458	0.827	1.914
bigg	0.863	0.819	0.817	0.67	0.813	1.082	0.805

Так же модель нужно было проверить модель на переобучение.
Первый способ — построить прогноз для какого-то участка исходной дороги. Результаты приведены на Рис. 17 и 18.
Второй предложенный способ — построить прогноз для обратной дороги Химки — Тосно. Результат показан на Рис. 19.
Видно, что данный метод склонен к переобучению на новых участках дорог, поэтому в дальнейших работах предлагается разработать методы настройки параметров данного алгоритма.

Прогноз алгоритма **DistrAlg** представлен на Рис. 20. По оси Y отложена вероятность попадания в интервал ширины $h = 600$с и $h = 300$с, а по оси X — середина каждого интервала.

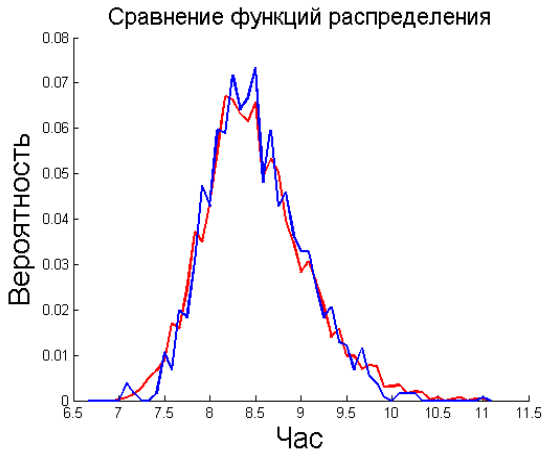

Рис. 12: Сравнение вероятности попадания значения в заданный интервал для прогноза (красным) и контрольной выборки для интервалов шириной 300с.

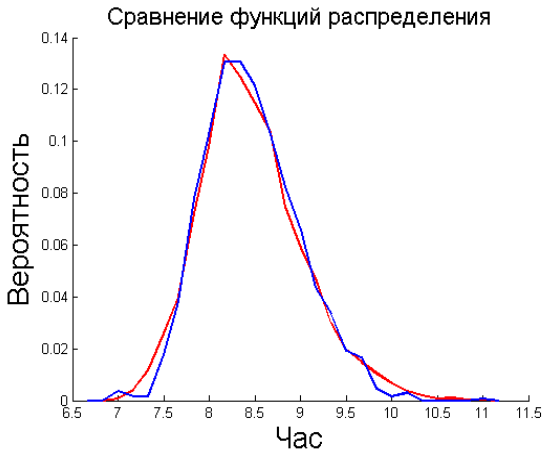

Рис. 13: Сравнение вероятности попадания значения в заданный интервал для прогноза (красным) и контрольной выборки для интервалов шириной 600с.

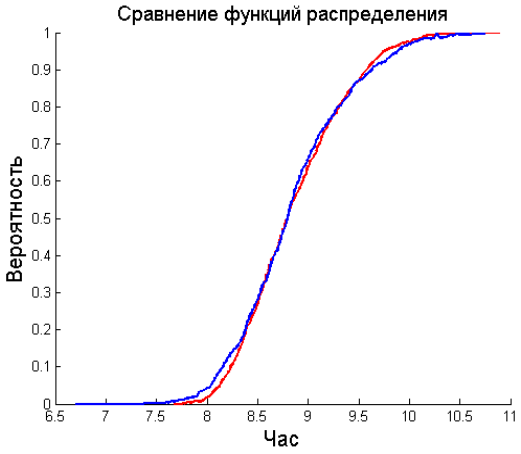

Рис. 14: Сравнение функции распределения контрольной выборки и прогноза (красным) при p_c=0.6

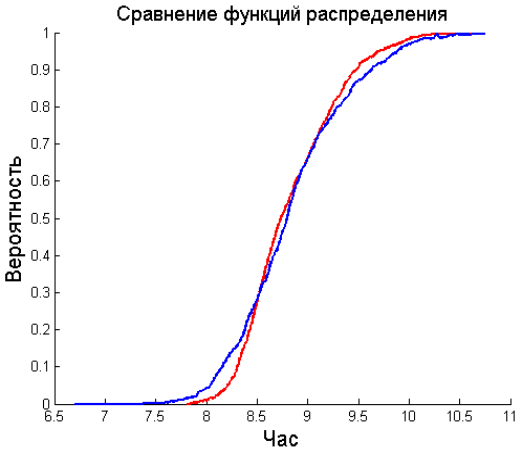

Рис. 15: Сравнение функции распределения контрольной выборки и прогноза (красным) при p_c=0.7

5 Заключение

Построено несколько алгоритмов, строящих прогноз времени прибытия:

- Алгоритм сложения средних времен проездов отрезков **BaseAlg**,
- Алгоритм сложения гистограмм **HistAlg**,
- Алгоритм построения эмпирической функции распределения методом сэмплирования **DistrAlg**.

Был сделан анализ качества каждого из них. Показано, что алгоритмы, основанные на сложении гистограмм и вычисления среднего времени проезда дают не очень точные результаты. В отличии от них, алгоритм **DistrAlg** показал хорошее качество прогноза, а так же обладает устойчивостью к многократным запускам и прореживанию историй проездов. Дальнейшие исследования будут направлены на улучшение обобщающих способностей алгоритма и разработку методов подбора параметров.

Список литературы

[1] F.M. Sanders C.P.I.J. van Hinsbergen, J.W.C. van Lint. Short term traffic prediction models. *ITS World Congress, Beijing, China*, October 2007.

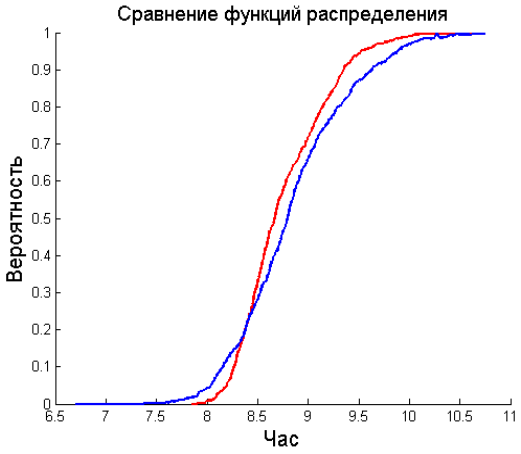

Рис. 16: Сравнение функции распределения контрольной выборки и прогноза (красным) при p_c=0.8

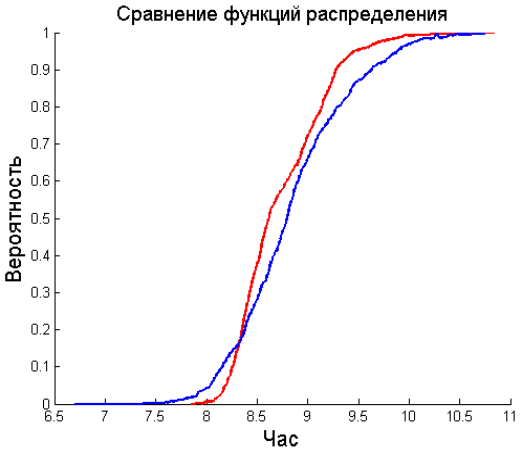

Рис. 17: Сравнение функции распределения контрольной выборки и прогноза (красным) при p_c=0.9

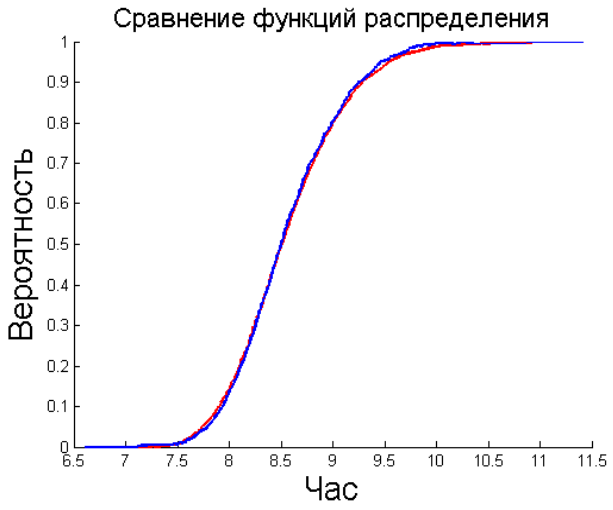

Рис. 18: Сравнение функций распределения прогноза (красным) и контрольной выборки.

[2] http://doc.utwente.nl/57639/1/thesis_Huisken.pdf.

[3] http://www.esafety-effects-database.org/applications_11.html.

[4] http://elar.urfu.ru/bitstream/10995/3060/1/russir-2010-06.pdf.

[5] http://www.csee.usf.edu/REU/publications/Persad%20Maharaj%20-%20PathPrediction%20-%20july%2031.pdf.

[6] http://www.siam.org/proceedings/datamining/2009/dm09_108_idet.pdf.

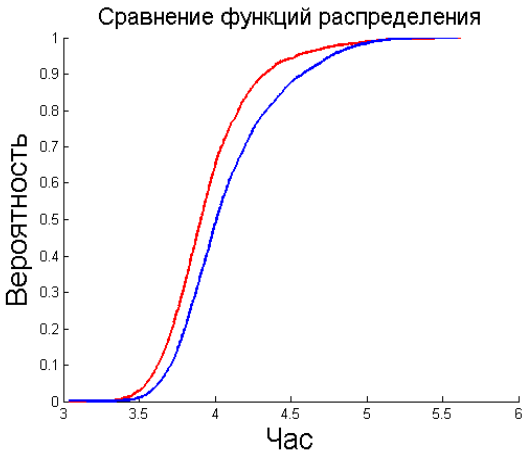

Рис. 19: Сравнение функций распределения прогноза (красным) и контрольной выборки для участка дороги между 150м и 400м отрезком.

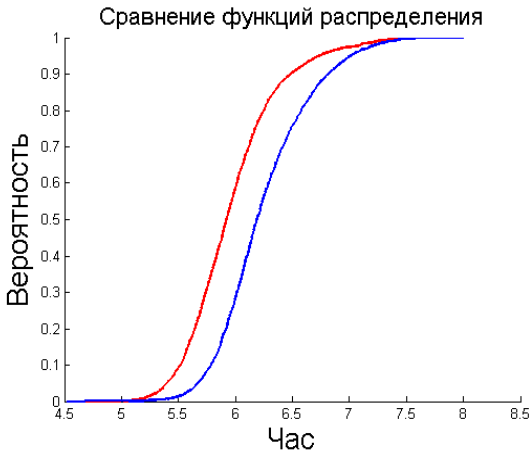

Рис. 20: Сравнение функций распределения прогноза (красным) и контрольной выборки для участка дороги между 0м и 300м отрезком.

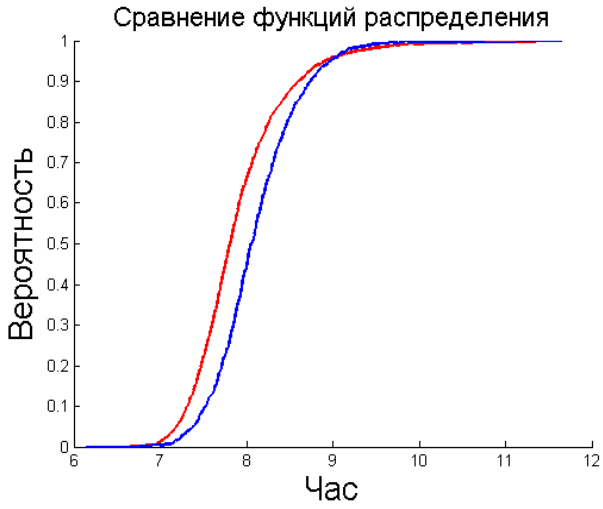

Рис. 21: Сравнение функций распределения прогноза (красным) и контрольной выборки для обратной дороги.

[7] http://en.wikipedia.org/wiki/Kernel_(statistics).

[8] http://ru.wikipedia.org/wiki/Разложение_Холецкого. Факторизация Холецкого.

[9] http://ssom.transportation.org/Documents/MwSWZDI-2001-Drakopoulos-TIPS.pdf.

[10] http://www.cost285.itu.edu.tr/tempodoc/TD05_12.pdf.

[11] R. (1980). Granger, C. W. J.; Joyeux. "an introduction to long-memory time series models and fractional differencing". *Journal of Time Series Analysis*, 1: 15–30.

[12] http://en.wikipedia.org/wiki/Autoregressive_fractionally_integrated_moving_average.

[13] http://cts.virginia.edu/docs/UVACTS-15-11-71.pdf.

[14] http://www.transport.iis.u-tokyo.ac.jp/publications/2004-017.pdf.

[15] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.202.3791&rep=rep1&type=pdf.

[16] http://company.yandex.ru/technologies/yaprobki/.

[17] http://www.integprog.ru/route2/.

[18] http://www.openstreetmap.org/.

[19] http://ru.wikipedia.org/wiki/Критерий_согласия_Колмогорова.