Прогнозирование времени прибытия автомобильных рейсов, совершающих междугородние перевозки

А. В. Шульга

Научный руководитель: Ю. В. Чехович
Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

24 июня 2013 г.

Цель работы

Дана история проездов автомобилей в виде таблицы треков H и граф дорог $\mathcal{G}$;

Необходимо спрогнозировать время прибытия T автомобильного рейса, совершающего междугороднюю перевозку по заданному набору ребер графа дорог, используя данные о проездах автомобилей в предыдущие моменты времени.

Основные определения

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 $\mathcal{G}$;

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.

Постановка задачи

Дано

множество отрезков заданного маршрута $\mathcal{M} : \{i\}, i = \overline{1, n}$;
история проездов каждого отрезка H_i;
Объект $\bar{\mathbf{x}}$;
T_0 — время старта;
Контрольная выборка $X^k = \{\bar{\mathbf{x}}_\mathbf{1}, \ldots, \bar{\mathbf{x}}_\mathbf{k}\}$ с ответами в виде времен проездов всего маршрута $Y^k = \{T_1, \ldots, T_k\}$.

Требуется построить функцию f()

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

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

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

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

Функционал качества

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

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

Задача минимизации

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

Этапы решения задачи

1. Обработка входных данных;
2. Выделение отрезков на графе дорог;
3. Построение истории проездов для каждого из отрезков;
4. Фильтрация истории проездов;
5. Построение алгоритмов (обучение, контроль);
6. Сравнение алгоритмов.

Алгоритм сложения средних времен проездов отрезков (далее BaseAlg)

Вход: входные параметры Θ_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 $\tilde{H}_i = H_i(\Theta_i)$;
3. Пересчитать входные параметры для следующего отрезка Θ_{i+1};
4. $T := T + t_i$;

Алгоритм сложения гистограмм (далее **HistAlg**)

Предполагается, что отрезки независимые.

Вход: входные параметры Θ_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 $\tilde{H}_i = H_i(\Theta_i)$;
3. **Выход из цикла**
4. Применить алгоритм сложения гистограмм соседних отрезков для всех отрезков с получением итоговой гистограммы G_T:
   $G_3 = G_1 + G_2$
   $p_3(t) = \sum\limits_{t_1+t_2=t} p_1(t_1) \cdot p_2(t_2)$
5. Построить прогноз по G_T

Алгоритм построения эмпирической функции распределения методом сэмплирования (далее DistrAlg)

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 пары соседних отрезков $(i-1, i)$ вычислить коэффициент корреляции c_i;
4. **Для всех** $j = 1, \ldots, N = 1000 \ldots 30000$
5. инициализировать $T_j := 0$
6. $r_1 = \mathrm{rand}(0, 1)$, $T_j := F_1^{-1}(r_1)$
7. **Для всех** $i = 2, \ldots, n$
8. $r_i = \mathrm{rand}(r_{i-1} - a, r_{i-1} + b)$

DistrAlg

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.

Вычислительный эксперимент. Предобработка данных

Таблица H не содержит разбиение треков на маршруты, поэтому их пришлось разбивать «вручную».

Из графа дорог брался только маршрут «Тосно-Химки».

В результате разбиения дороги получилось 505 отрезков, длиной от 2м до 10км.

Примененялась медианная фильтрация для историй проездов H_i из-за неточности GPS.

Контрольная выборка формировалась из тех маршрутов, которые проходили полностью от Тосно до Химки. Так же применялась фильтрация остановок. В итоге было получено 595 прецедентов.

Вычислительный эксперимент. Предварительная фильтрация данных

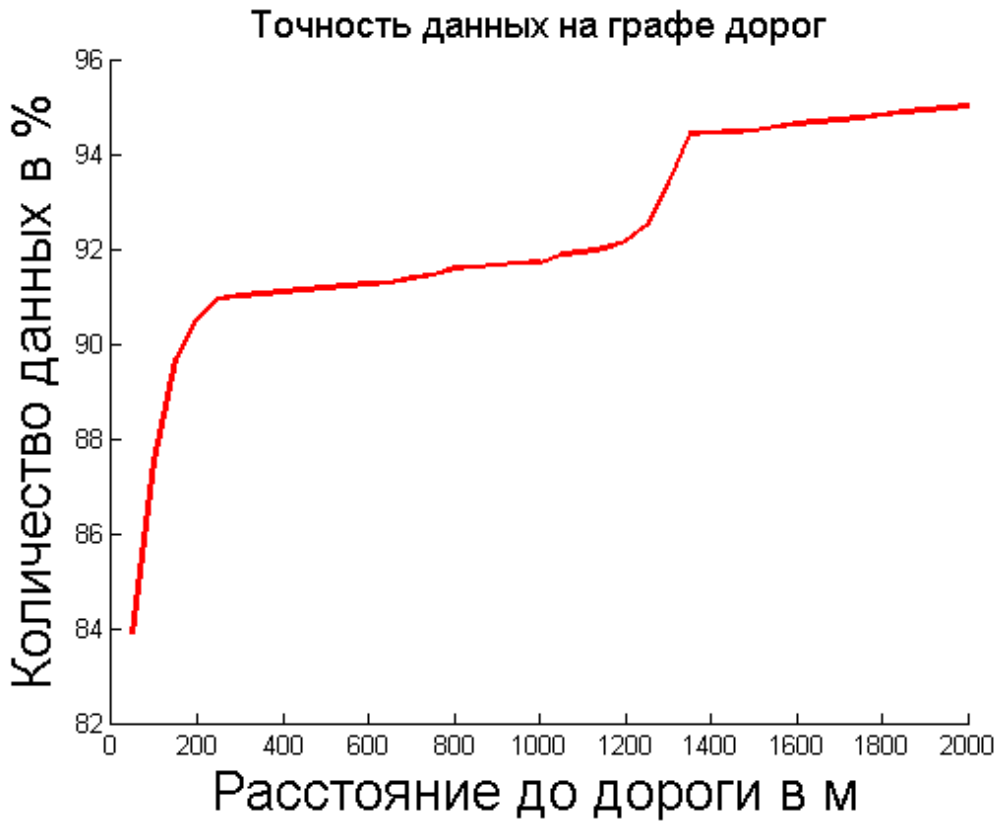

Рис. : Доля точек внутри полосы, длиной h, в зависимости от h.

Видно, что 90 % данных лежит в 200м полосе

Вычислительный эксперимент. Алгоритм сложения средних времен проездов отрезков

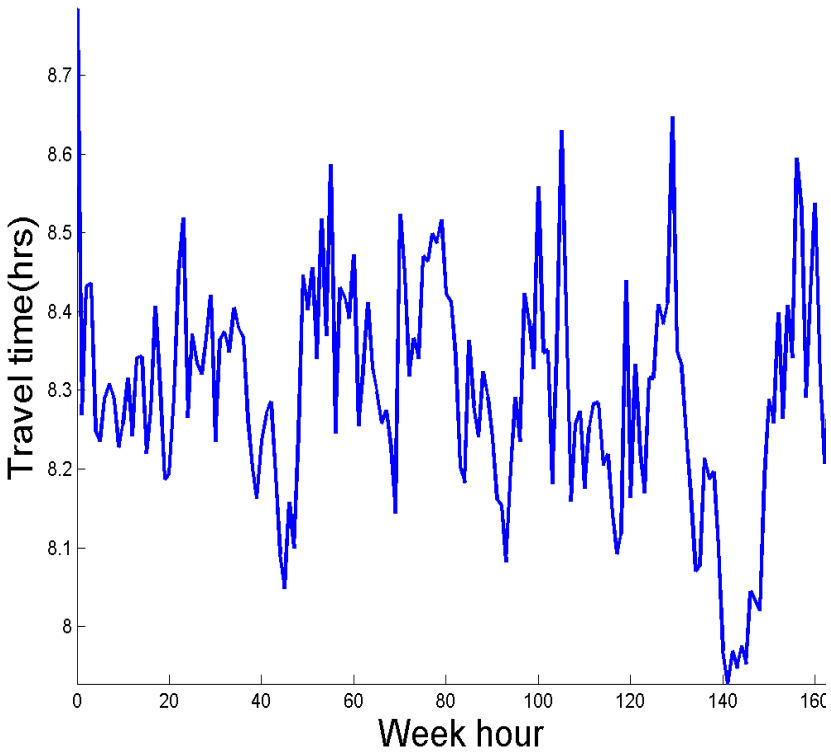

Рис. : Прогноз алгоритма **BaseAlg** в зависимости от часа недели.

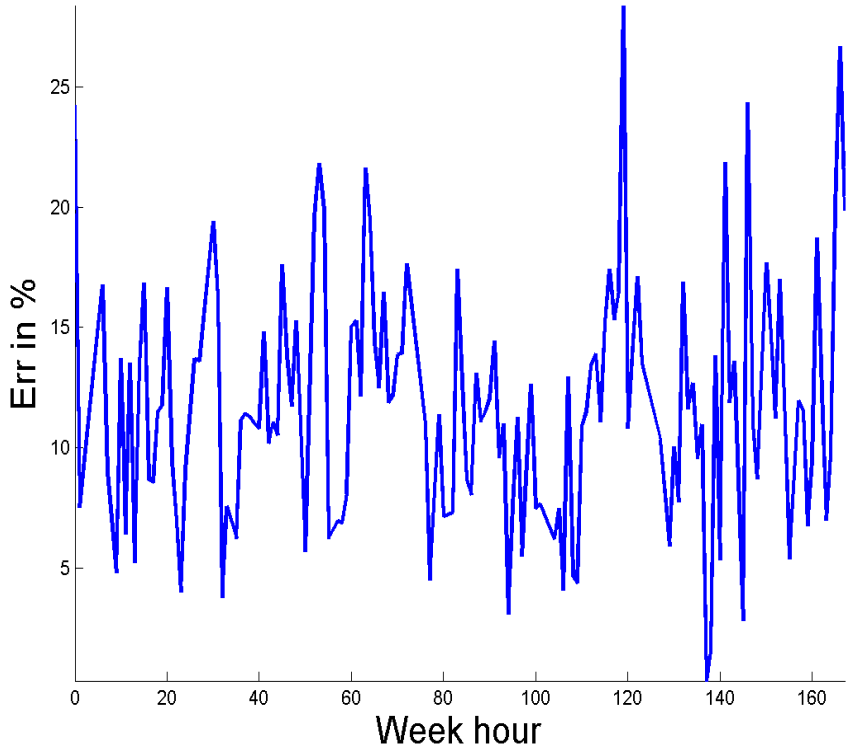

Рис. : Ошибка алгоритма **BaseAlg** в сравнении с контрольной выборкой в зависимости от часа недели.

Вычислительный эксперимент. Алгоритм сложения гистограмм

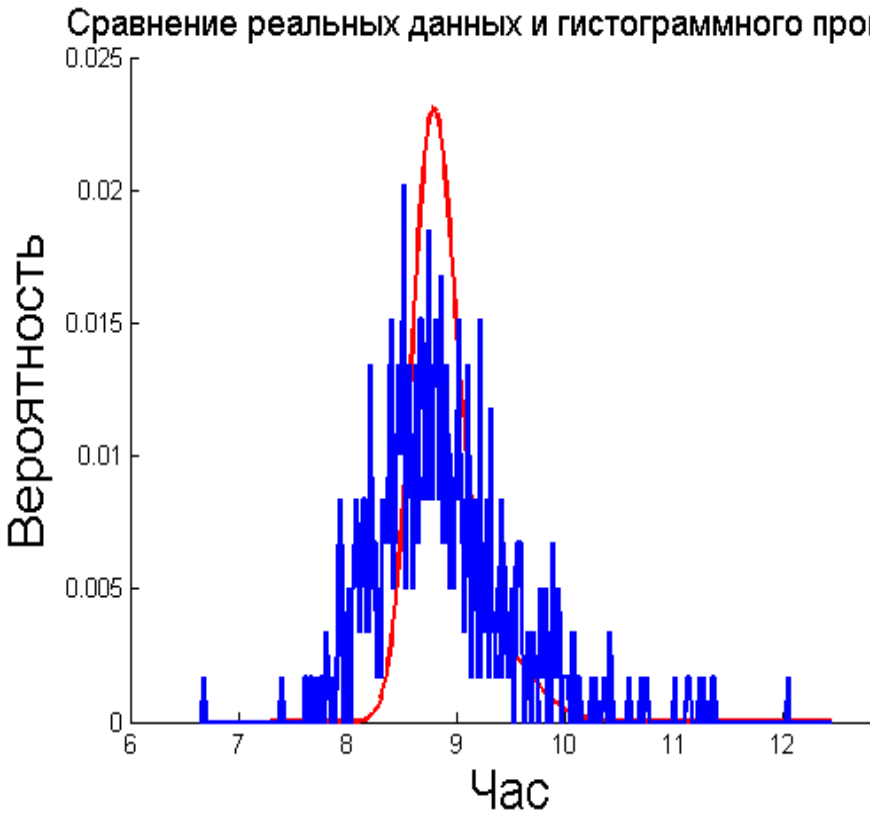

Рис. : Итоговая гистограмма для **HistAlg** (красным) в сравнении с контрольной выборкой, $h = 5$с.

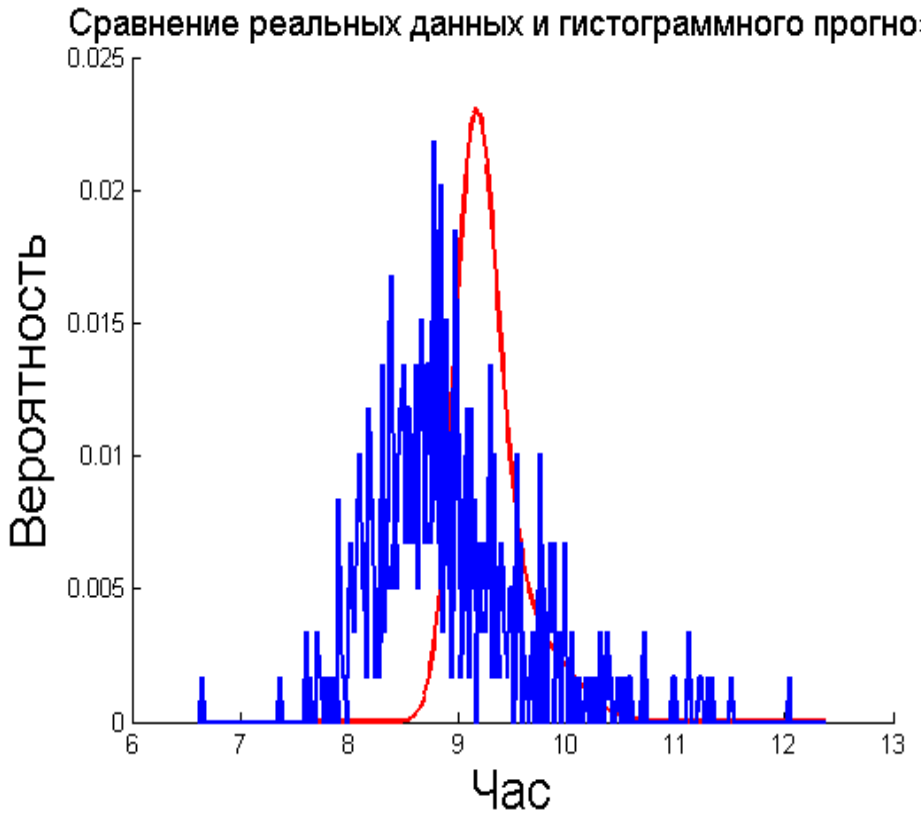

Рис. : Итоговая гистограмма для **HistAlg** (красным) в сравнении с контрольной выборкой, $h = 10$с.

Вычислительный эксперимент. Алгоритм построения эмпирической функции распределения методом сэмплирования

В результате подбора параметров эмпирическим методом, был установлен вид функций φ_1, φ_2:

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

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

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

Вычислительный эксперимент. Алгоритм построения эмпирической функции распределения методом сэмплирования

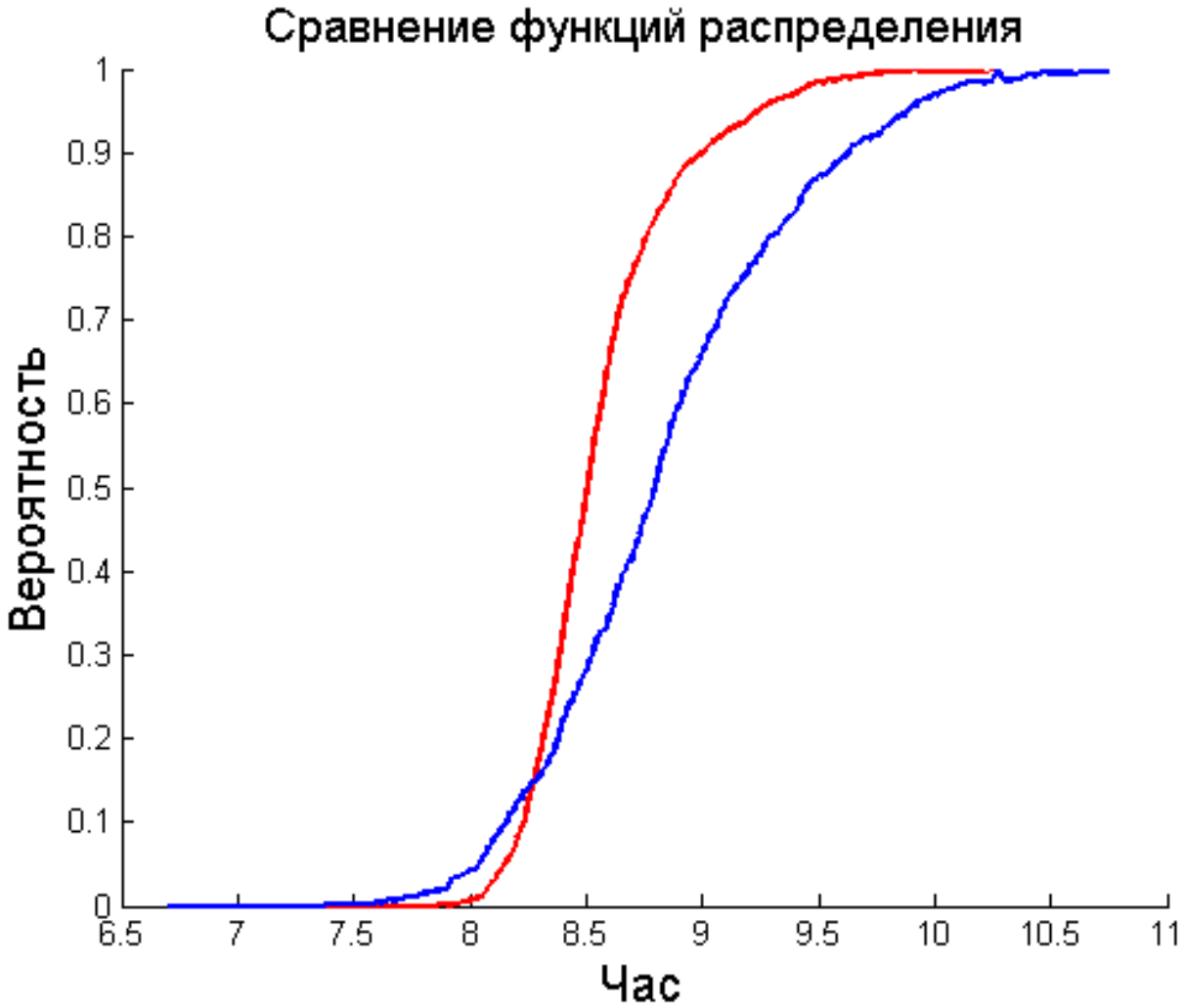

Рис. : Сравнение функции распределения прогноза (красным) с контрольной выборкой при гипотезе о независимости отрезков.

Вычислительный эксперимент. Алгоритм построения эмпирической функции распределения методом сэмплирования

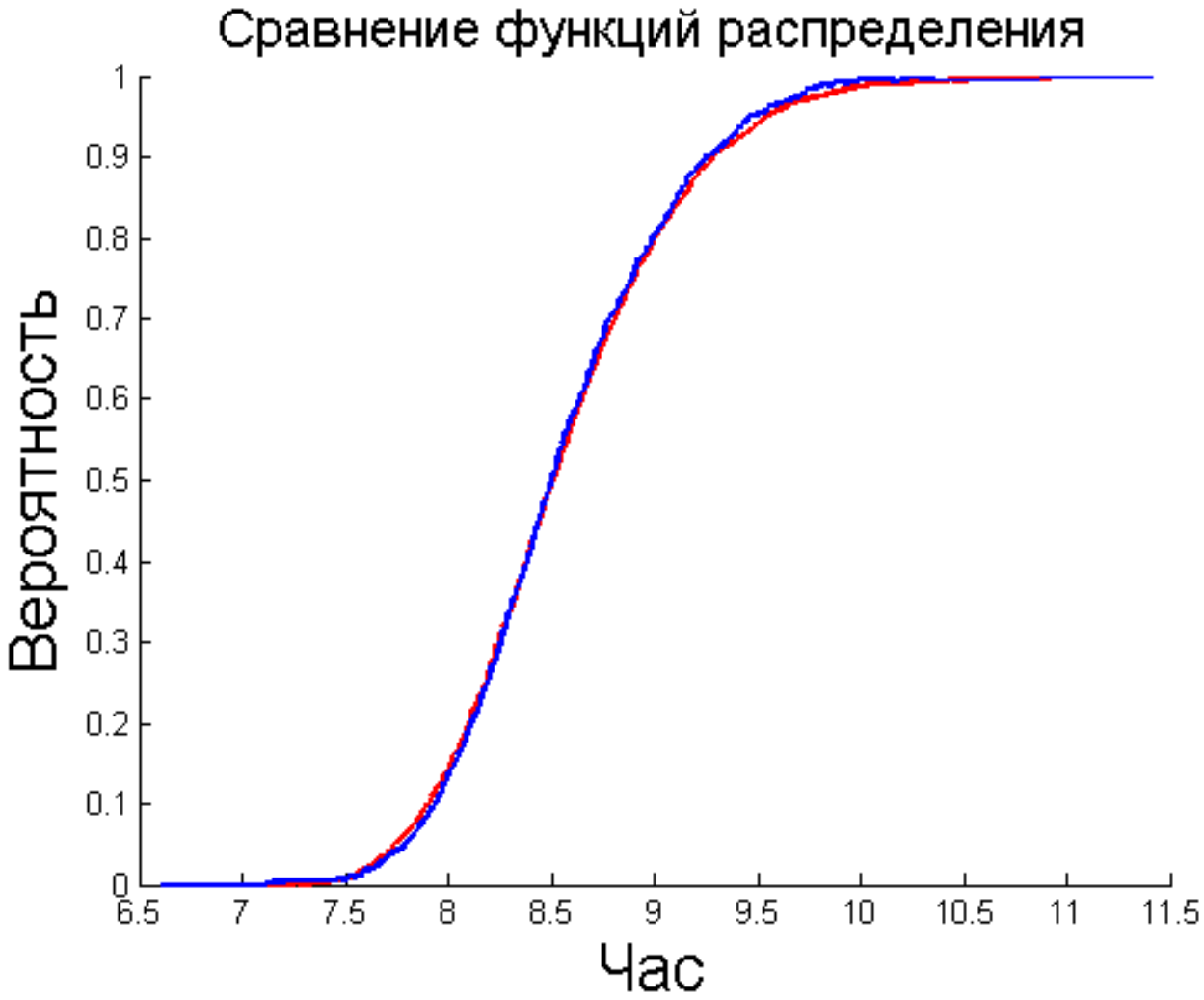

Рис. : Сравнение функций распределения прогноза (красным) и контрольной выборки.

Вычислительный эксперимент. Алгоритм построения эмпирической функции распределения методом сэмплирования. Устойчивость

Таблица : 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

Таблица : Устойчивость модели при многократном запуске сэмплирований

№	1	2	3	4	5	6	7	8
λ_α	0.805	0.985	0.805	1.202	0.598	0.989	0.859	0.797

Вычислительный эксперимент. Алгоритм построения эмпирической функции распределения методом сэмплирования. Устойчивость

Таблица : Устойчивость модели при разреживании 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

Вычислительный эксперимент. Алгоритм построения эмпирической функции распределения методом сэмплирования. Переобучение

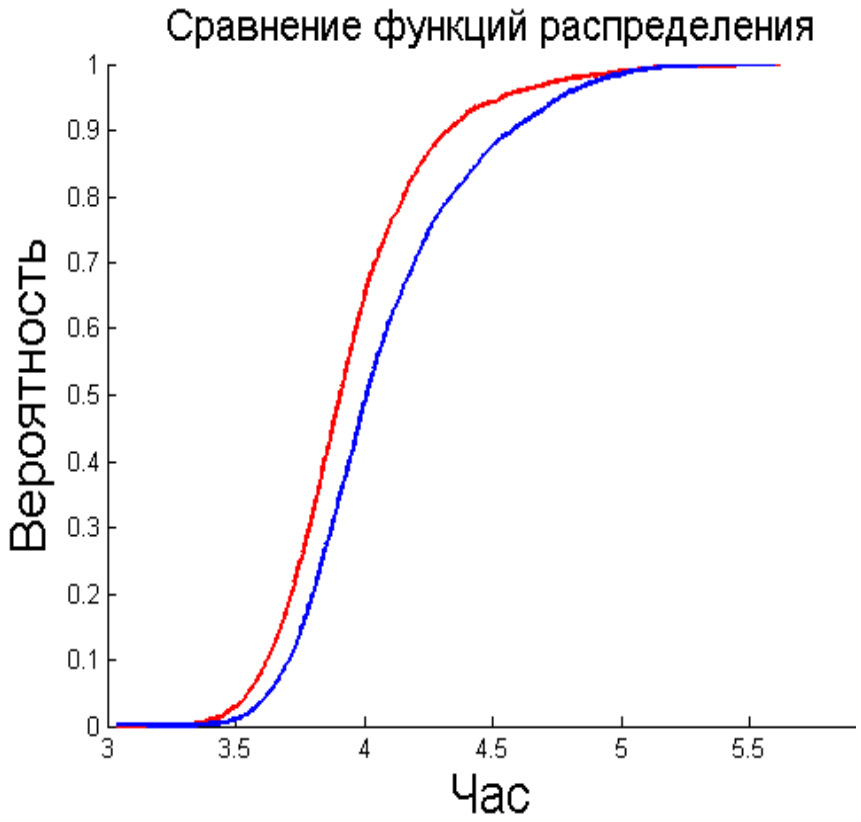

Рис. : Сравнение функций распределения для участка дороги между 150 и 400 отрезком.

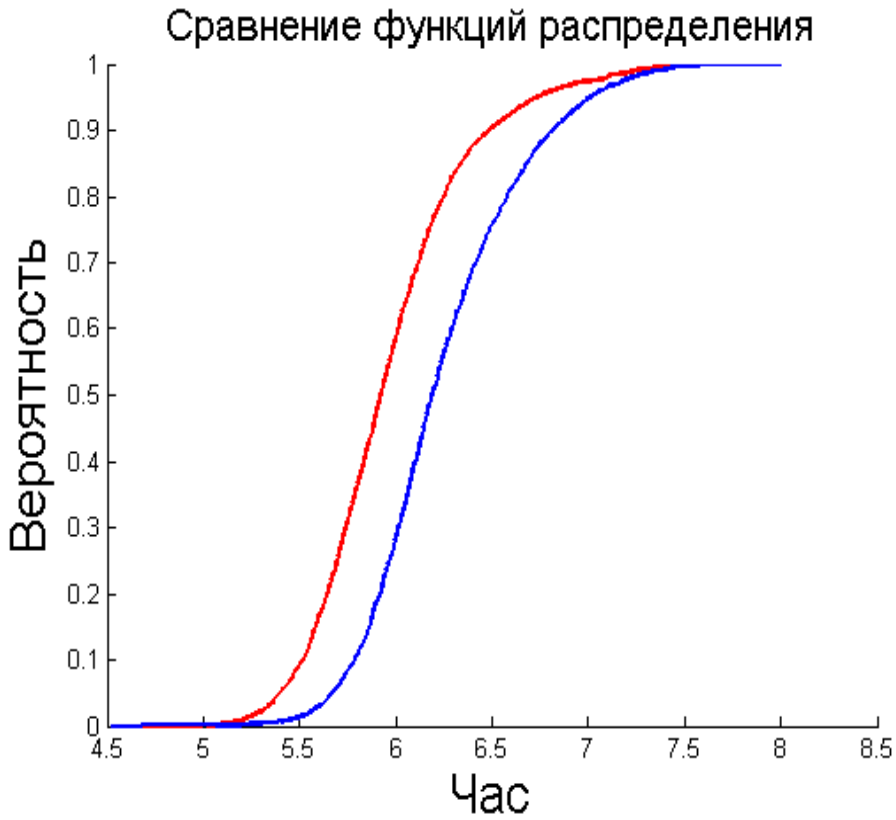

Рис. : Сравнение функций распределения для участка дороги между 0 и 300 отрезком.

Вычислительный эксперимент. Алгоритм построения эмпирической функции распределения методом сэмплирования. Переобучение

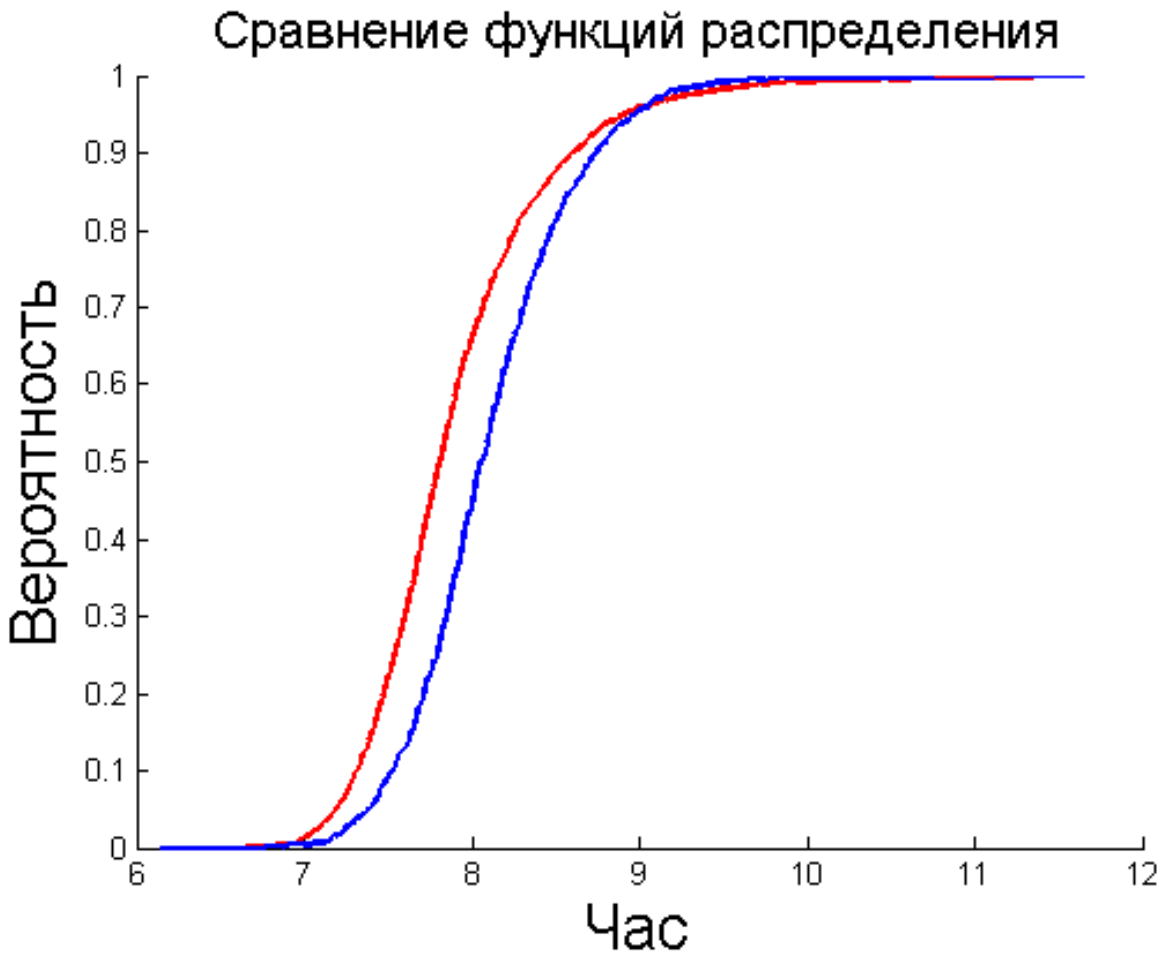

Рис. : Прогноз (красным) и контрольная выборка для обратной дороги из Химок в Тосно.

Вычислительный эксперимент. Алгоритм построения эмпирической функции распределения методом сэмплирования. Результаты прогноза

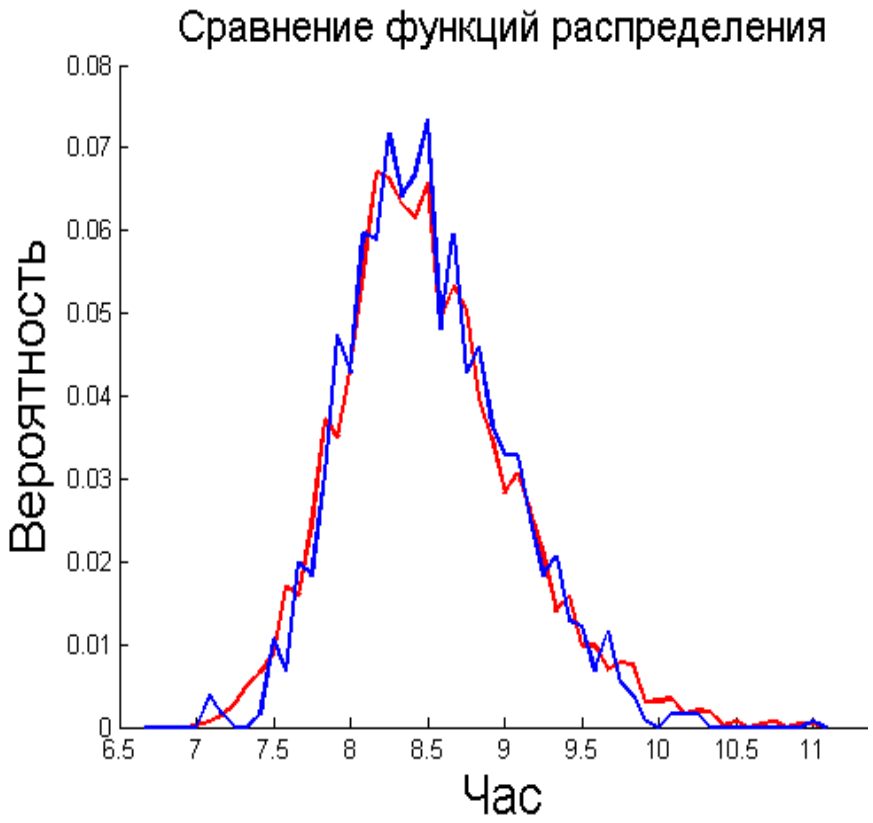

Рис. : Сравнение вероятности попадания значения в заданный интервал шириной 300с.

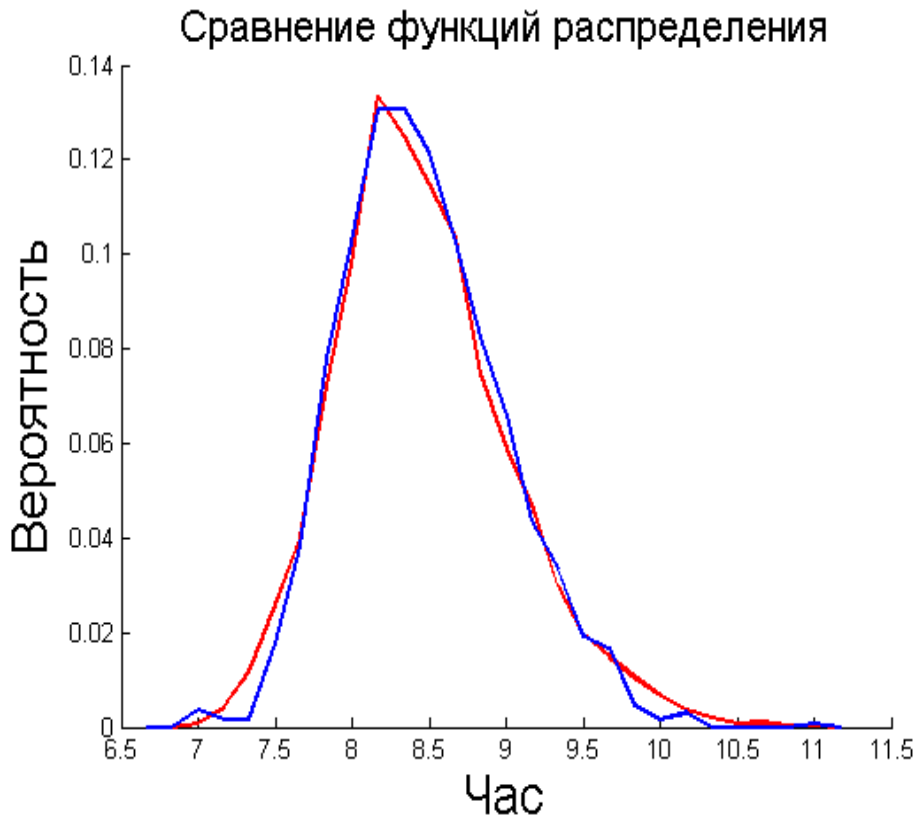

Рис. : Сравнение вероятности попадания значения в заданный интервал шириной 600с.

Заключение

Результаты: 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.