Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математической статистики

Горбунов Сергей Алексеевич

# Статистический анализ данных трафика виртуального мобильного оператора

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

**Научный руководитель:**
д.ф.-м.н., доцент
А. К. Горшенин

Москва, 2022

# Оглавление

# Введение

Интернет-трафик является приоритетным способом обмена данными для большинства пользователей мобильных устройств. Исследование структуры статистических параметров объёмов отправленного и полученного трафика позволяет оценить нагрузку на цифровую сеть. В данной работе основной акцент будет сделан на задаче прогнозирования различных характеристик объёмов трафика и задаче выявления аномально больших объёмов трафика с использованием обобщённых гамма-распределений. Целью такого анализа является получение необходимой информации для проектирования и оптимизации сети, а также совершенствования технологий обеспечения качества обслуживания [1].

Актуальность задачи прогнозирования характеристик сети мобильной связи обусловлена необходимостью оценки возможных сценариев поведения сети при изменении параметров эксплуатации. Согласно исследованию [2], к причинам подобного анализа относятся: введение в эксплуатацию нового оборудования каналов связи, изменение маршрутизации трафика, введение в сервис новых мультимедийных услуг и дополнительных сервисов. Следовательно, необходим математический аппарат, позволяющий на основе данных, собранных автоматизированными измерительными комплексами, выполнить прогнозирование характеристик сети.

Информация об аномальных значениях объёмов трафика и нетипичном поведении пользователей зачастую используется для выявления атак и вторжений [3]. Статистический подход к обнаружению аномалий пользуется популярностью в задачах анализа мобильного трафика [4, 5, 6].

Анализируемые данные имеют две уникальные особенности: во-первых, они изначально представлены в «агрегированном» виде – объёмы трафика суммируются за час по каждому пользователю; во-вторых, каждое наблюдение имеет метку типа приложения, которое инициировало обмен данными. Первая особенность приводит к необходимости выбора вероятностной модели для аппроксимации распределений выборок объёмов просуммированного трафика за час. Для этой цели в работе предлагается использование обобщённого гамма-распределения [7].

# Глава 1

# Данные виртуального мобильного оператора

В качестве модели обслуживания трафика мобильного оператора рассматривается сота, в зоне действия которой находятся подвижные устройства. Каждое устройство имеет свой номер мобильного абонента цифровой сети с интеграцией служб MSISDN. Всего насчитывается 94 973 уникальных устройств.

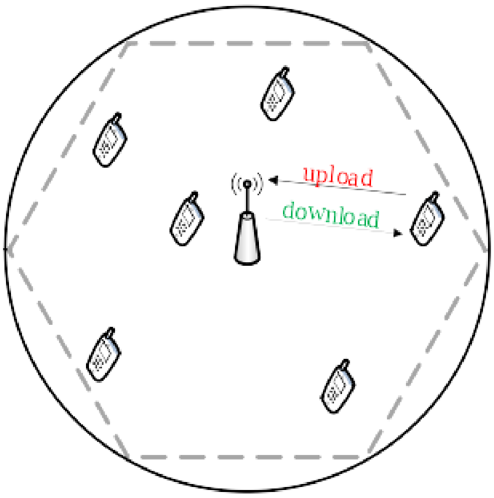



Рис. 1.1. Схема соты

Данные собирались в течение двух временных промежутков: 10.02.2018 – 22.02.2018 и 01.03.2018 — 04.03.2018. Каждый час по каждому активному устройству суммируется отправленный и полученный трафик в битах. Величины объёмов переданного и загруженного трафика имеют метки одного из типов приложений, инициирующих обмен данными. В таблице 1.1 количество бит поделено на $2^{20}$ для удобства дальнейшего анализа.

| | START_HOUR | MASKED_MSISDN |
|---|---|---|
| 1 | 2018-02-12 10:00:00 | 9BFA3001DE8C58C2453B69CE2E4A3704 |
| 2 | 2018-02-12 08:00:00 | 0C2C88351A593CD02727A6207EB85E9E |
| 3 | 2018-02-12 10:00:00 | B9F45E1542162096408D94DD94499B89 |
| 4 | 2018-02-12 10:00:00 | 3CF6B81BC186E4C188D69E1FE8919BB6 |
| 5 | 2018-02-11 18:00:00 | 7EE1FCE60945D869A14EF30E896E9131 |
| 6 | 2018-02-11 17:00:00 | A746CCDCAC507B95C83A3475C63C6BFD |

| | APP_CLASS | UPLOAD | DOWNLOAD |
|---|---|---|---|
| 1 | Web Applications | 0.10618114 | 0.07458496 |
| 2 | Instant Messaging Applications | 0.01139259 | 0.00642872 |
| 3 | Web Applications | 64.03334808 | 4.03468609 |
| 4 | Web Applications | 0.01797104 | 0.00931644 |
| 5 | Streaming Applications | 16.42163754 | 0.38467503 |
| 6 | Web Applications | 0.22454739 | 0.12561035 |

Таблица 1.1. Данные мобильного оператора

- START_HOUR — дата и час, за который суммируется трафик;
- MASKED_MSISDN — идентификатор устройства пользователя;
- APP_CLASS — тип приложения, инициирующего обмен данными;
- UPLOAD — количество отправленных бит;

- DOWNLOAD — количество полученных бит.

Особенностью этих данных является почасовая агрегированность трафика: за каждый час в данных присутствует целая выборка просуммированного по каждому активному пользователю и приложению трафика. Всего выделяется 16 классов приложений:

1. `DB Transactions` — транзакции в базе данных (например, перевод средств с банковской карты через мобильное приложение);
2. `File Systems` — работа с удаленными файловыми системами;
3. `File Transfer` — передача файлов по протоколу FTP;
4. `Games` — трафик в онлайн-играх;
5. `Instant Messaging Applications` — системы мгновенного обмена сообщениями (мессенджеры WhatsApp, Viber, Telegram и другие);
6. `Legacy Protocols` — устаревшие протоколы;
7. `Mail` — электронная почта;
8. `Music Streaming` — потоковые сервисы для прослушивания музыки;
9. `Network Operation` — сетевые службы;
10. `Others` — прочие приложения;
11. `P2P Applications` — приложения, передача данных в которых основана на принципах одноранговых сетей (например, приложения, работающие по протоколу Bittorrent);
12. `Security` — данные онлайн-видеокамер, данные с датчиков сигнализации;

13. `Streaming Applications` — потоковые сервисы для просмотра фильмов и видео-чаты;

14. `Terminals` — мобильные терминалы для оплаты банковскими картами;

15. `VoIP` — IP-телефония (например, звонки через WhatsApp или Skype);

16. `Web Applications` — клиент-серверные приложения, в которых клиент взаимодействует с сервером при помощи браузера (например, Microsoft Office Online, Google Documents).

Подробнее изучить, что «скрывается» за метками типов приложений, можно, например, в статьях [8, 9].

Данные содержат 44 009 843 наблюдений, однако эти наблюдения распределены по типам приложений неравномерно. В таблице 1.2 отчетливо виден дисбаланс классов приложений по количеству наблюдений в них.

| Тип приложений | Число наблюдений |
|---|---:|
| Web Applications | 10 238 001 |
| Others | 6 857 095 |
| Instant Messaging Applications | 5 878 799 |
| Games | 5 512 957 |
| File Transfer | 4 520 415 |
| Mail | 2 787 705 |
| Streaming Applications | 2 562 286 |
| VoIP | 1 931 893 |
| Security | 1 769 702 |
| Music Streaming | 837 507 |
| Network Operation | 675 477 |
| P2P Applications | 291 558 |
| Terminals | 129 417 |
| File Systems | 10 852 |
| DB Transactions | 6 156 |
| Legacy Protocols | 23 |

Таблица 1.2. Количество наблюдений по типам приложений

# Глава 2

# Обобщённое гамма-распределение в задачах анализа мобильного трафика

## 2.1 Две параметризации плотности распределения

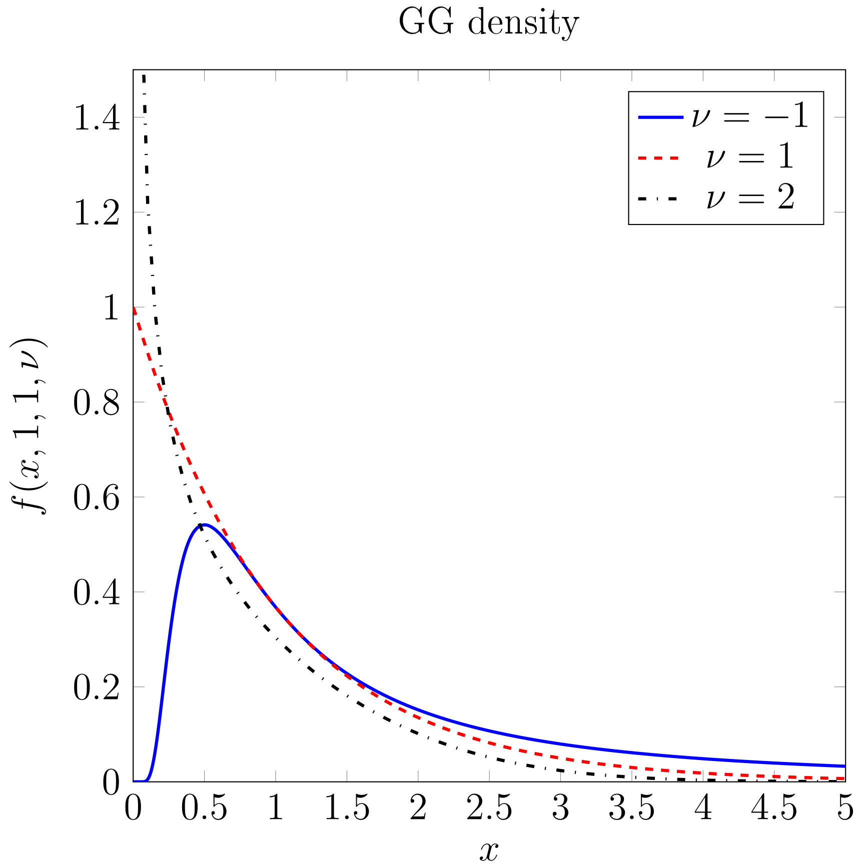
Для аппроксимации распределений объёмов полученного и отправленого трафика будет использована параметризация обобщённого гамма-распределения с плотностью (2.1).

$$f(x;\mu,\sigma,\nu) = \frac{|\nu|\theta^{\theta}z^{\theta}e^{-\theta z}}{\Gamma(\theta)x},\ x > 0 \tag{2.1}$$

$$\mu > 0,\ \sigma > 0,\ -\infty < \nu < \infty,\ \nu \neq 0,$$
$$z = \left(\frac{x}{\mu}\right)^{\nu},\ \theta = \frac{1}{\sigma^2\nu^2},$$

где $\mu$ — параметр расположения (location), $\sigma$ — параметр масштаба (scale), $\nu$ — параметр формы (shape).

На графике ниже представлены формы плотностей обобщённого гамма-распределения для $\mu = 1,\ \sigma = 1,\ \nu = -1, 1, 2$.




Частные случаи обобщённого гамма-распределения в этой параметризации:

- GG$(\mu, \sigma, 1)$=GA$(\mu, \sigma)$;
- GG$(\mu, \sigma^{-1}, \sigma)$=WEI$(\mu, \sigma)$.

Для статистической процедуры выявления аномальных наблюдений, описанной в разделе 2.5 будет использована параметризация обобщённого гамма-распределения с плотностью (2.2).

$$g(x; r, \gamma, \mu_1) = \frac{|\gamma|\mu_1^r}{\Gamma(r)} x^{\gamma r - 1} e^{-\mu_1 x^\gamma},\ x > 0 \tag{2.2}$$

$$r > 0,\ \mu_1 > 0,\ -\infty < \gamma < \infty,\ \gamma \neq 0$$

**Предложение 1.** *Для того чтобы перейти от параметризации обобщенного гамма-распределения с плотностью (2.1) к параметризации с плотностью (2.2), необходимо выполнить следующую замену параметров:* $r = \frac{1}{\sigma^2\nu^2},\ \gamma = \nu,\ \mu_1 = \frac{1}{\sigma^2\nu^2\mu^\nu}$.

*Доказательство.* Пусть $x > 0$, напомним, что в плотности с параметризацией (2.1) $z = \left(\frac{x}{\mu}\right)^\nu,\ \theta = \frac{1}{\sigma^2\nu^2}$, откуда следует

$$f(x;\mu,\sigma,\nu) = \frac{|\nu|\theta^\theta z^\theta e^{-\theta z}}{\Gamma(\theta)x} = \frac{|\nu|\left(\frac{1}{\sigma^2\nu^2}\right)^{\frac{1}{\sigma^2\nu^2}}\left(\frac{x}{\mu}\right)^{\frac{\nu}{\sigma^2\nu^2}} e^{-\frac{1}{\sigma^2\nu^2}\left(\frac{x}{\mu}\right)^\nu}}{\Gamma(\frac{1}{\sigma^2\nu^2})x} =$$

$$= \frac{|\nu|\left(\frac{1}{\sigma^2\nu^2\mu^\nu}\right)^{\frac{1}{\sigma^2\nu^2}} x^{\frac{1}{\sigma^2\nu^2}\nu-1} e^{-\frac{1}{\sigma^2\nu^2\mu^\nu}x^\nu}}{\Gamma(\frac{1}{\sigma^2\nu^2})} = \left\{\begin{aligned} & r = \frac{1}{\sigma^2\nu^2} \\ & \gamma = \nu \\ & \mu_1 = \frac{1}{\sigma^2\nu^2\mu^\nu} \end{aligned}\right\} =$$

$$= \frac{|\gamma|\mu_1^r}{\Gamma(r)} x^{\gamma r-1} e^{-\mu_1 x^\gamma} = g(x;r,\gamma,\mu_1)$$

□

Семейство обобщённых гамма-распределений $g(x;r,\gamma,\mu)$ содержит практически все самые популярные абсолютно непрерывные распределения, сосредоточенные на положительной полупрямой. В частности, это семейство содержит:

- Гамма-распределение ($\gamma = 1$) и его частные случаи
  - Показательное распределение ($\gamma = 1,\ r = 1$),
  - Распределение Эрланга ($\gamma = 1,\ r \in \mathbb{N}$),
  - Распределение хи-квадрат ($\gamma = 1,\ \mu = \frac{1}{2}$);
- Распределение Накагами ($\gamma = 2$);
- Полунормальное распределение (распределение максимального значения стандартного винеровского процесса на $[0,1]$) ($\gamma = 2,\ r = \frac{1}{2}$);

- Распределение Рэлея ($\gamma = 2,\ r = 1$);
- Хи-распределение ($\gamma = 2,\ \mu = \frac{1}{\sqrt{2}}$);
- Распределение Максвелла (распределение абсолютных значений скоростей молекул в разреженном газе) ($\gamma = 2,\ r = \frac{3}{2}$);
- Распределение Вейбулла-Гнеденко ($r = 1,\ \gamma > 0$);
- Сложенное экспоненциальное степенное распределение ($\gamma > 0$, $r = \frac{1}{\gamma}$);
- Обратное гамма-распределение ($\gamma = -1$) и его частный случай
  - Распределение Леви ($\gamma = -1,\ r = \frac{1}{2}$);
- Распределение Фреше ($r = 1,\ \gamma < 0$)

и прочие законы распределения. Предельным законом для семейства обобщённых гамма-распределений является

- Логнормальное распределение ($r \to \infty$).

## 2.2 Аппроксимация эмпирических распределений объёма трафика

В ходе исследования было установлено, что распределения объёмов отправленного и полученного трафика за различные временные промежутки, а также по различным типам приложений, хорошо аппроксимируются обобщённым гамма-распределением. На рисунках 2.1 – 2.4 приведены гистограммы выборок, полученных агрегацией наблюдений по различным временным окнам, типам трафика и приложениям, а также информация о статистическом качестве аппроксимации. Синей линией на рисунках проведены кривые плотности соответствующих выборкам обобщённых гамма-распределений, параметры которых оценивались методом максимального правдоподобия функцией `fitdist` из библиотеки `fitdistrplus` на языке `R`.

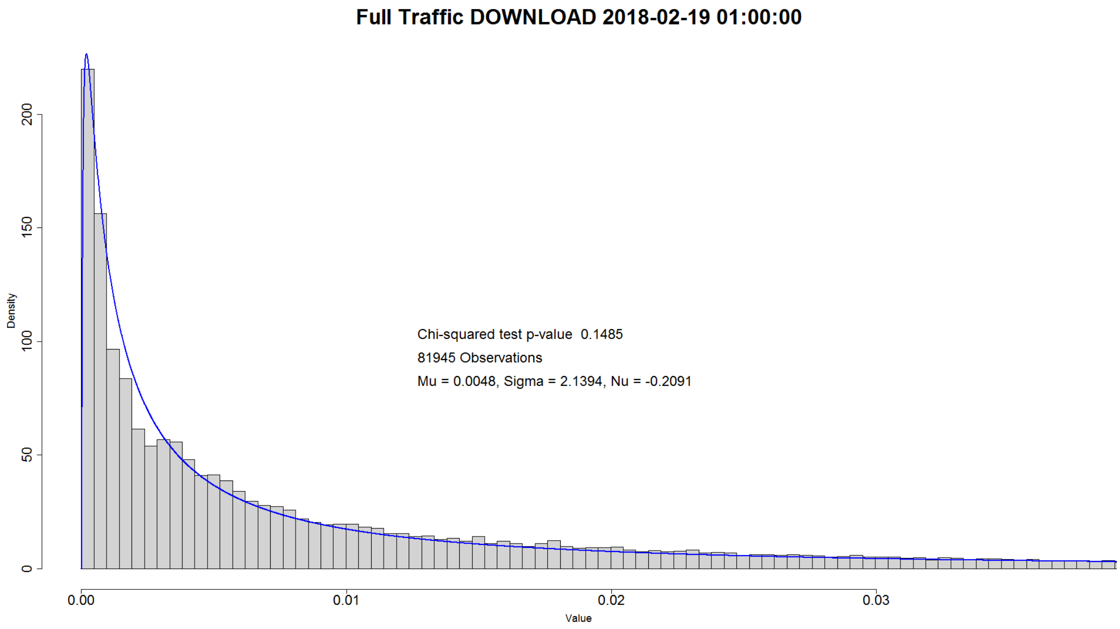



Рис. 2.1. Выборка всего полученного трафика за один час одного дня

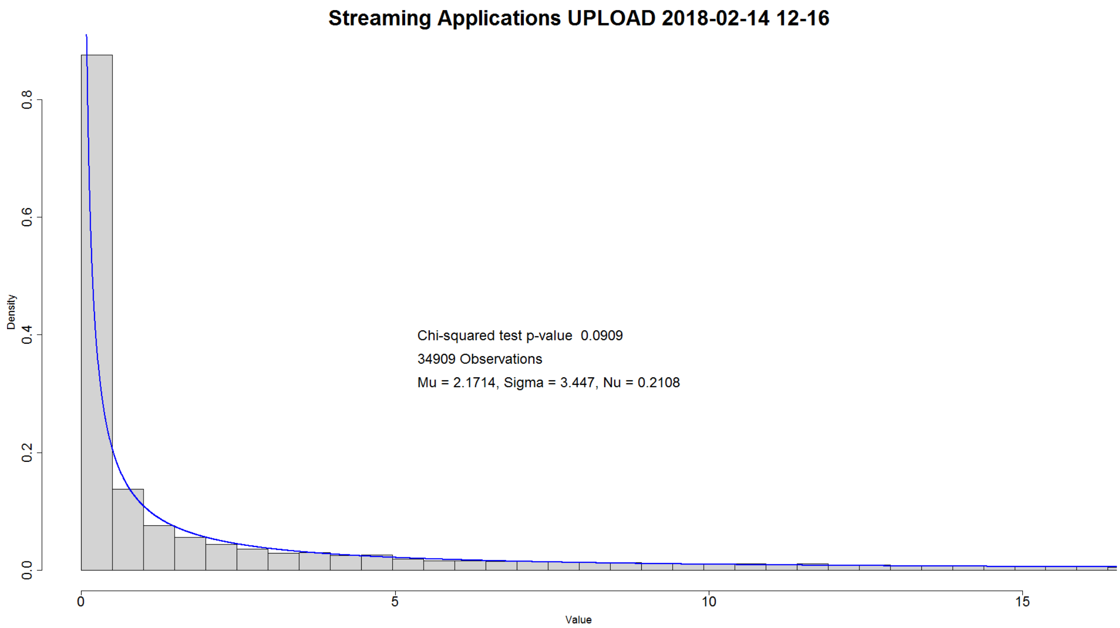



Рис. 2.2. Выборка отправленного трафика по приложению `Streaming Applications` за четыре часа одного дня

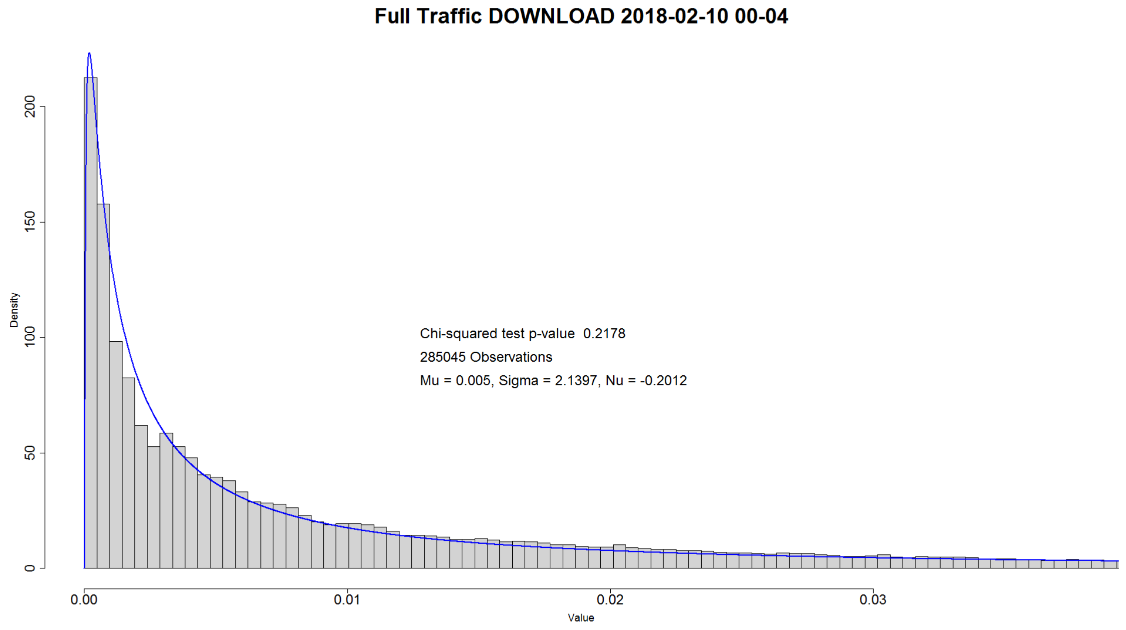



Рис. 2.3. Выборка всего полученного трафика за четыре часа одного дня

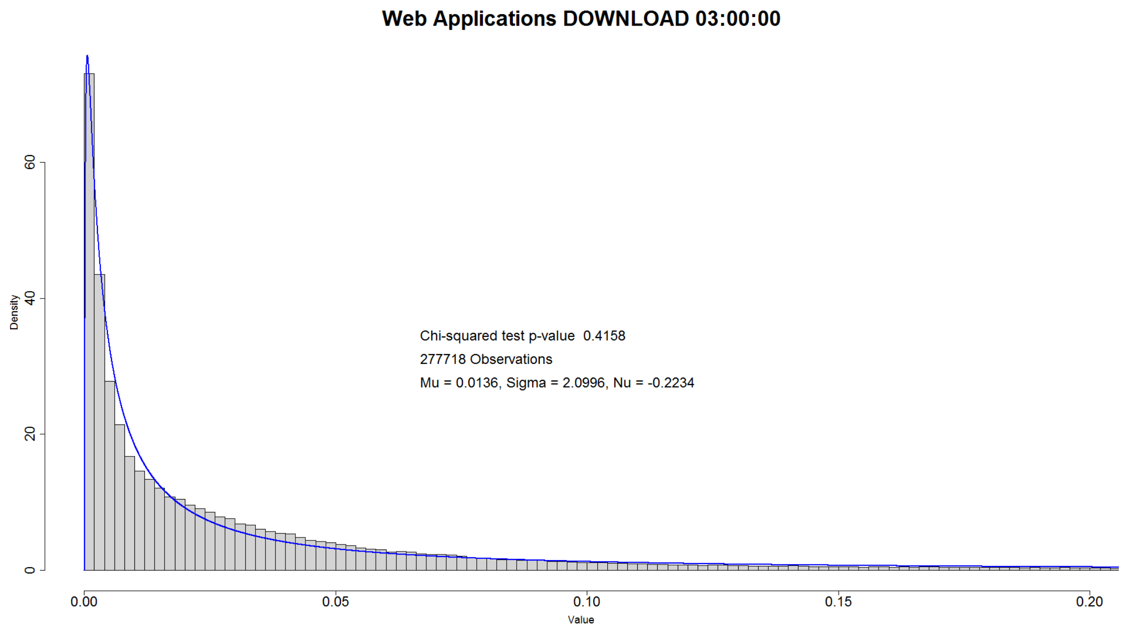



Рис. 2.4. Выборка полученного трафика по приложению `Web Applications` за один час по всем дням

В таблице 2.1 приведены значения параметров обобщённого гамма-распределения, оценённые по выборкам объёмов отправленного и полученного трафика, агрегированным с одночасовым окном без разбиения по приложениям, и p-значения теста хи-квадрат [10], округлённые до второго знака после запятой.

| Дата:час | Отправленный трафик | | | | Полученный трафик | | | |
|---|---|---|---|---|---|---|---|---|
| | $\mu$ | $\sigma$ | $\nu$ | p-значение | $\mu$ | $\sigma$ | $\nu$ | p-значение |
| 2018-02-10:01 | 0.0057 | 2.465 | -0.335 | 1 | 0.0052 | 2.226 | -0.207 | 0.64 |
| 2018-02-10:02 | 0.0055 | 2.351 | -0.317 | 1 | 0.0052 | 2.126 | -0.188 | 0.03 |
| 2018-02-10:03 | 0.0050 | 2.287 | -0.297 | 0.94 | 0.0046 | 2.051 | -0.197 | 0.05 |
| 2018-02-10:04 | 0.0046 | 2.226 | -0.291 | 0.43 | 0.0042 | 1.986 | -0.204 | 0 |
| 2018-02-10:05 | 0.0044 | 2.221 | -0.274 | 0.67 | 0.0040 | 1.948 | -0.204 | 0 |
| 2018-02-10:06 | 0.0048 | 2.250 | -0.271 | 0.92 | 0.0043 | 1.976 | -0.180 | 0.02 |
| 2018-02-10:07 | 0.0056 | 2.349 | -0.289 | 1 | 0.0053 | 2.093 | -0.158 | 0 |
| 2018-02-10:08 | 0.0068 | 2.483 | -0.299 | 1 | 0.0062 | 2.205 | -0.159 | 0.45 |
| 2018-02-10:09 | 0.0065 | 2.612 | -0.303 | 1 | 0.0057 | 2.322 | -0.178 | 0.94 |
| 2018-02-10:10 | 0.0064 | 2.691 | -0.300 | 1 | 0.0056 | 2.391 | -0.179 | 0.91 |
| 2018-02-10:11 | 0.0063 | 2.730 | -0.298 | 1 | 0.0057 | 2.429 | -0.177 | 0.93 |
| 2018-02-10:12 | 0.0063 | 2.741 | -0.305 | 1 | 0.0060 | 2.464 | -0.169 | 0.88 |
| 2018-02-10:13 | 0.0064 | 2.762 | -0.308 | 1 | 0.0061 | 2.484 | -0.170 | 0.66 |
| 2018-02-10:14 | 0.0060 | 2.752 | -0.321 | 0.45 | 0.0060 | 2.487 | -0.174 | 0.11 |
| 2018-02-10:15 | 0.0061 | 2.764 | -0.314 | 0.46 | 0.0057 | 2.498 | -0.183 | 0.43 |
| 2018-02-10:16 | 0.0060 | 2.778 | -0.317 | 0.76 | 0.0057 | 2.494 | -0.189 | 0.31 |
| 2018-02-10:17 | 0.0062 | 2.791 | -0.316 | 0 | 0.0058 | 2.514 | -0.185 | 0.04 |
| 2018-02-10:18 | 0.0063 | 2.787 | -0.315 | 0.99 | 0.0059 | 2.515 | -0.182 | 0.5 |
| 2018-02-10:19 | 0.0066 | 2.809 | -0.311 | 0.19 | 0.0062 | 2.517 | -0.177 | 0.57 |
| 2018-02-10:20 | 0.0068 | 2.817 | -0.313 | 0 | 0.0063 | 2.528 | -0.182 | 0.82 |
| 2018-02-10:21 | 0.0065 | 2.847 | -0.318 | 0 | 0.0059 | 2.545 | -0.193 | 0.05 |
| 2018-02-10:22 | 0.0062 | 2.839 | -0.330 | 0 | 0.0059 | 2.563 | -0.195 | 0 |
| 2018-02-10:23 | 0.0054 | 2.764 | -0.346 | 0.1 | 0.0051 | 2.501 | -0.213 | 0 |
| 2018-02-11:00 | 0.0052 | 2.670 | -0.345 | 0.87 | 0.0048 | 2.402 | -0.220 | 0.08 |

Таблица 2.1. Параметры обобщённого гамма-распределения и p-значения теста хи-квадрат для выборок, полученных агрегацией с одночасовым окном, за первые сутки в данных

## 2.3 Распределение объёмов трафика в течение суток, кластеризация приложений

Оценим параметры обобщённого гамма-распределения для каждого приложения, кроме `Legacy Protocols` (в силу малочисленности это класса (см. таблицу 1.2) в дальнейшем он будет исключен из анализа), за каждый час в сутках. Выборки будем формировать, учитывая только час, за который трафик был просуммирован, подобно тому, как это сделано на рисунке 2.4. Предложенный метод позволит нам анализировать изменения параметров $\mu$, $\sigma$, $\nu$ в течение суток для каждого приложения. На рисунке 2.5 продемонстрировано изменение параметров для приложения `Security`.

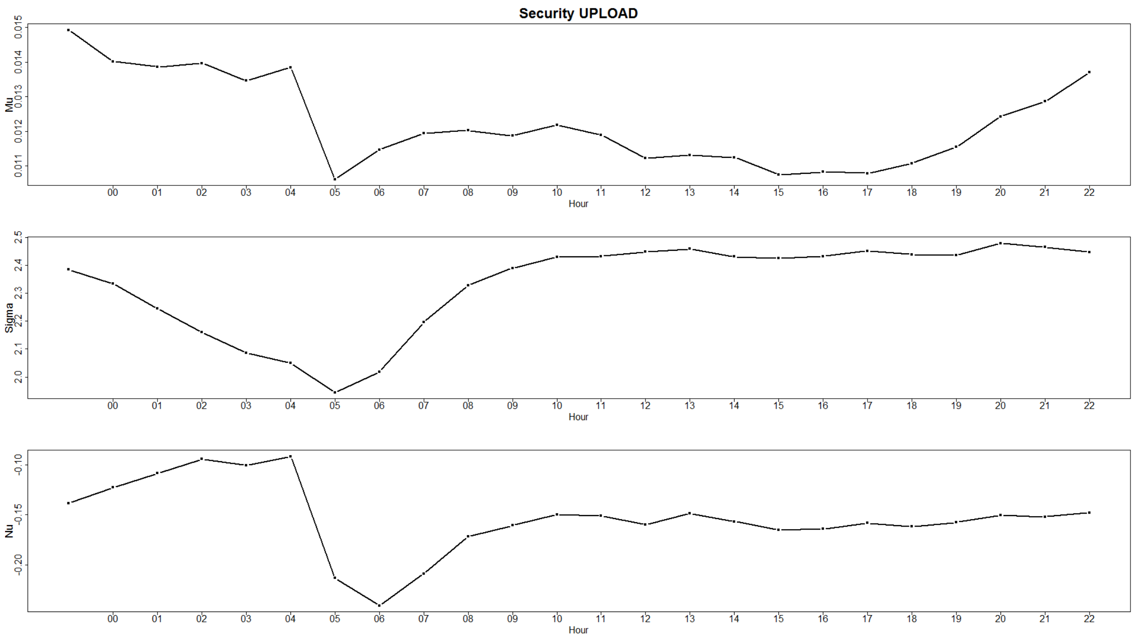



Рис. 2.5. Изменение параметров распределения объёмов отправленного трафика для приложения `Secutiry` в течение суток

Больший интерес представляет анализ таких графиков сразу для всех приложений. Изменение параметров распределений каждого приложения в течение дня представлено на рискунках 2.7 – 2.12, а цветовая легенда приведена на рисунке 2.6.

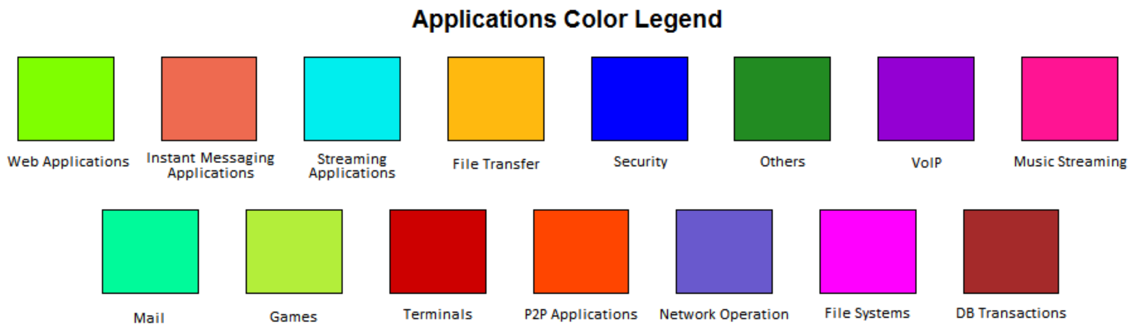



Рис. 2.6. Цветовая легенда

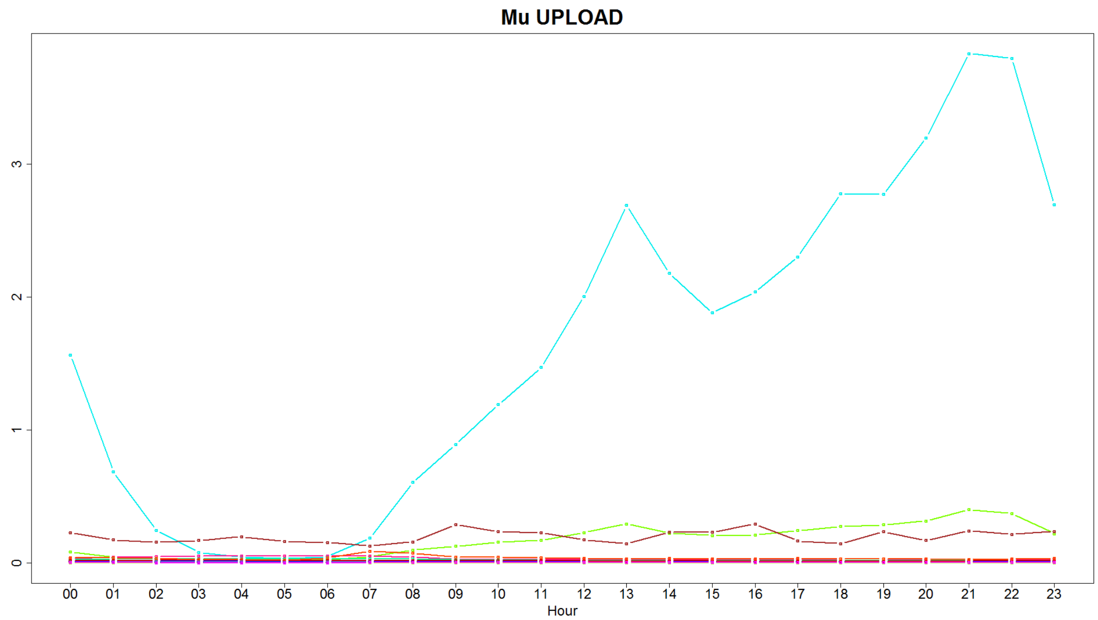



Рис. 2.7. Параметр расположения для отправленного трафика

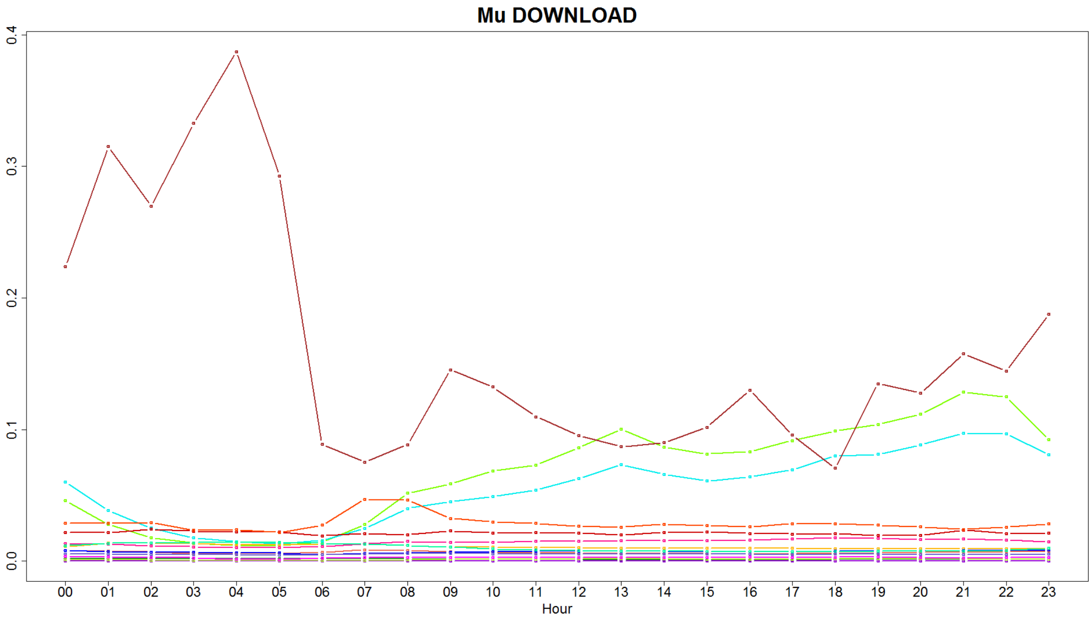



Рис. 2.8. Параметр расположения для полученного трафика

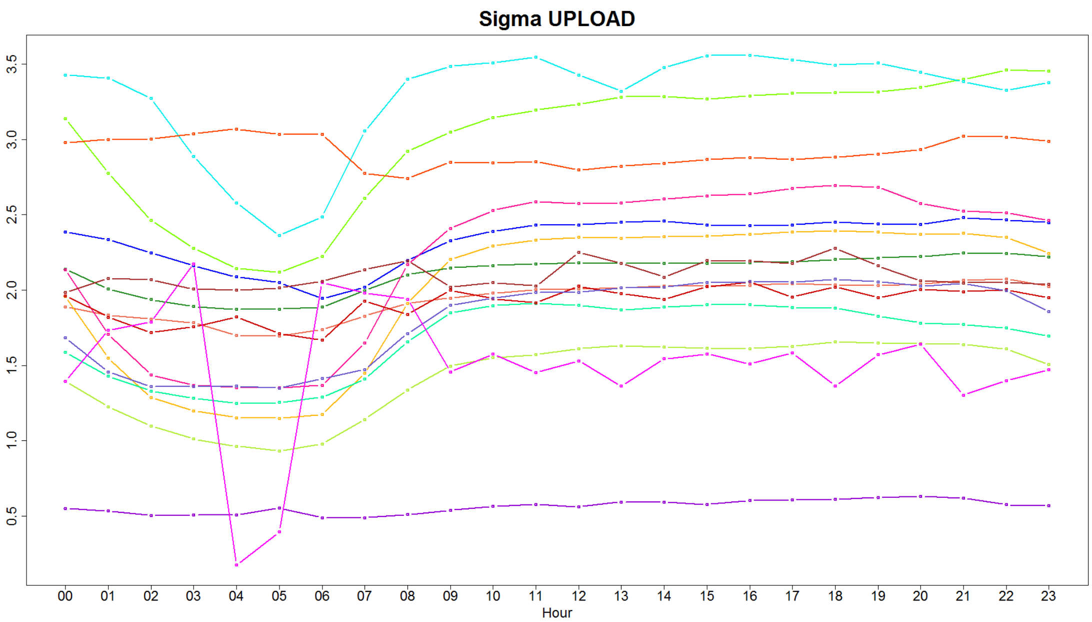



Рис. 2.9. Параметр масштаба для отправленного трафика

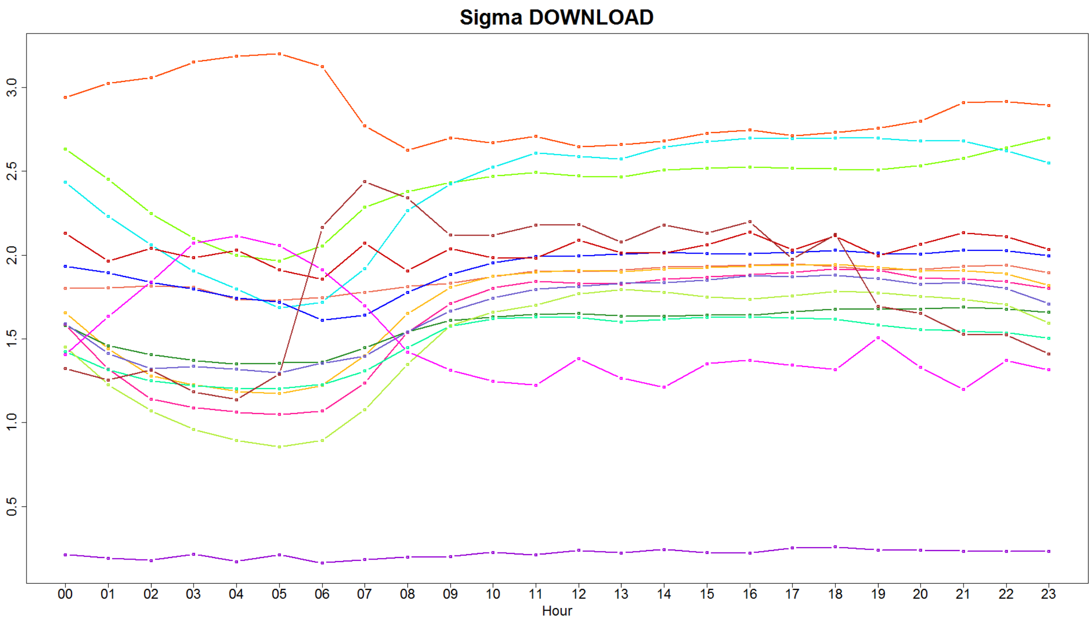



Рис. 2.10. Параметр масштаба для полученного трафика

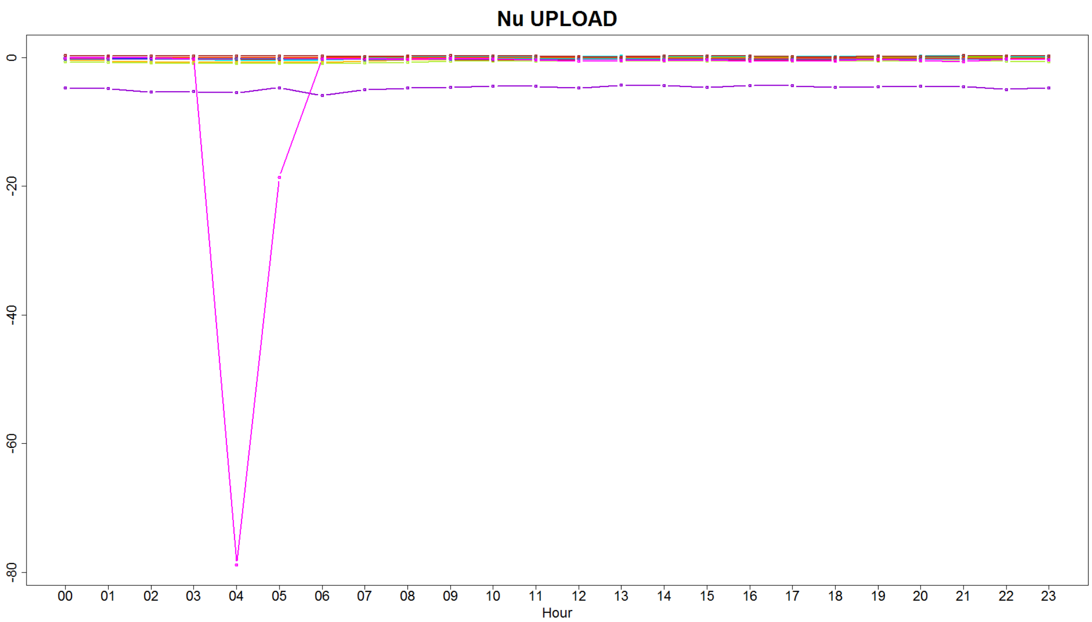



Рис. 2.11. Параметр формы для отправленного трафика

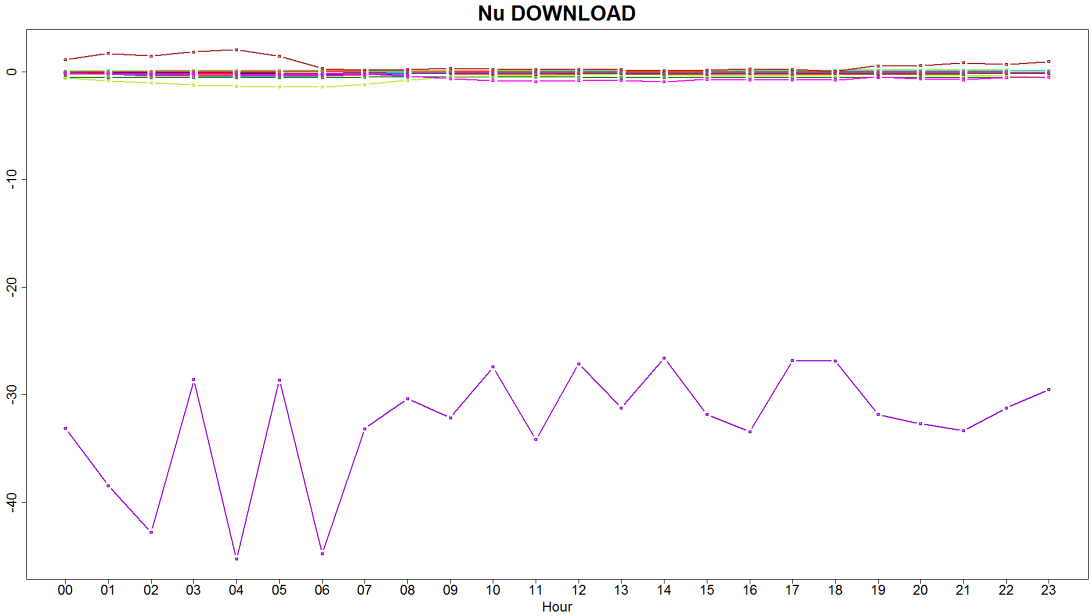



Рис. 2.12. Параметр формы для полученного трафика

Легко видеть, что параметры распределений для некоторых приложений ведут себя похожим образом в течение суток, а параметры распределений для приложения `VoIP` зачастую проходят обособленно. Отсюда приходит идея кластеризовать метки приложений так, чтобы приложения с похожими распределениями в течение суток попали в один кластер. Для этого используем матрицу с параметрами распределений за каждый час как признаковое пространство для приложений и проведем иерархическую кластеризацию функцией `hclust` из библиотеки `stats` на языке `R`. Подсчет расстояний между кластерами будет производиться методом Уорда [11]. Дендрограмма, полученная в результате иерархической кластеризации, представлена на рисунке 2.13.

На дендрограмме отчетливо видны три кластера, притом приложение `VoIP` ожидаемо отделилось от всех других приложений. Данная кластеризация может быть полезна при прогнозировании параметров распределений приложений (см. раздел 2.4). Дело в том, что прогнозировать параметры распределений малочисленных приложений с маленьким окном по времени может быть затруднительно, так

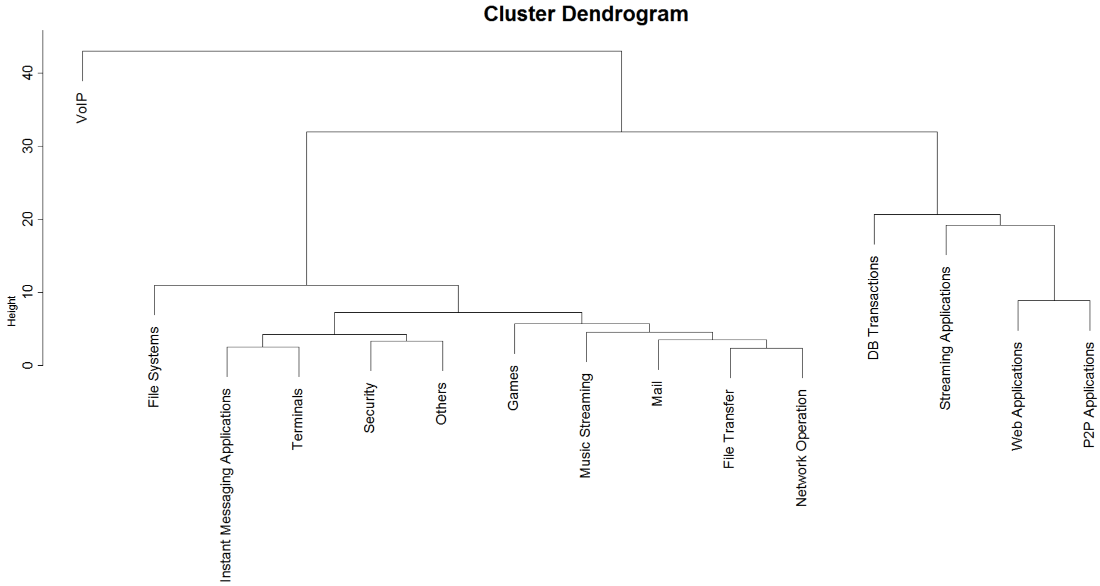



Рис. 2.13. Дендрограмма иерархической кластеризации приложений

как сформированные выборки будут слишком маленьких объёмов. Вместо этого можно прогнозировать параметры распределений не для одного приложения, а для целого кластера схожих по распределению объёмов трафика в течение дня приложений.

## 2.4 Прогнозирование параметров распределения объёмов полного трафика

В этом разделе основное внимание будет уделено прогнозированию параметров распределения объёмов полного (без разбиения на приложения) трафика с одночасовым окном, однако такая же техника прогнозирования применима и при других агрегациях трафика (выборе другого окна по времени и приложениям: анализ может проводиться как для полного трафика, так и для каждого приложения по отдельности, возможен также анализ кластеров приложений, полученных в разделе 2.3).

Будем проходить окном в один час (при работе с этими данными это наименьшее возможное окно, так как трафик суммируется по часам) по первому интервалу времени, за который данные собирались, формировать выборки и оценивать параметры распределения объёмов полного трафика. Таким образом, для каждого типа трафика получаем три временных ряда параметров распределения (см. таблицу 2.1), к ним добавим количество наблюдений в выборке, количество уникальных пользователей, суммарный объём трафика и средний объём трафика, приходящийся на уникального пользователя, за час. Итого имеем семь временных рядов для каждого из двух типов трафика.

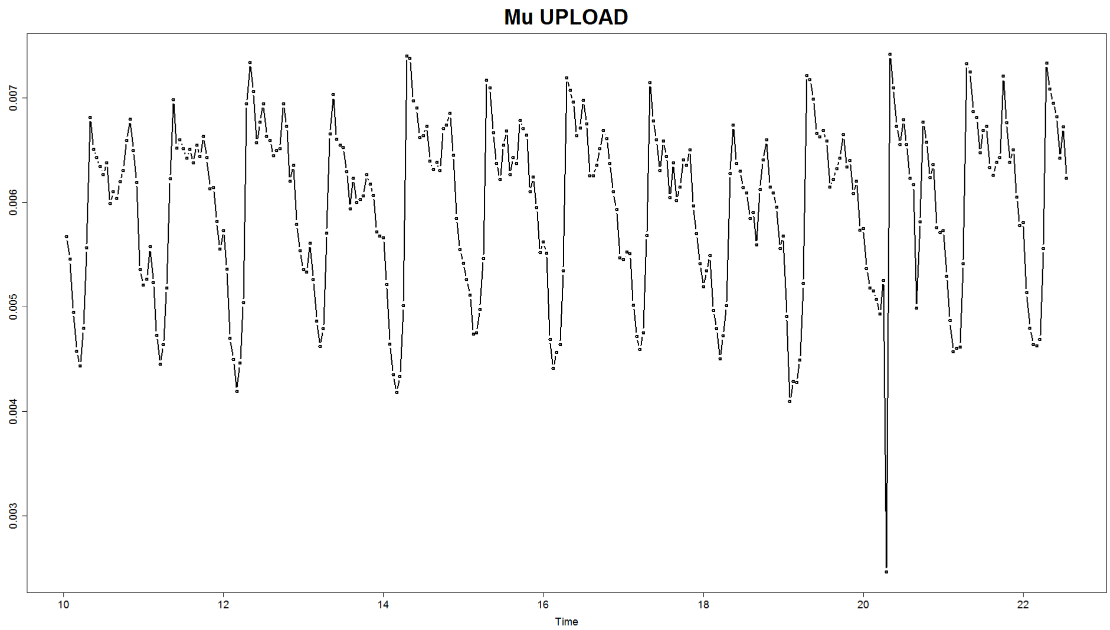



Рис. 2.14. Временной ряд `Mu Upload`

На рисунке 2.14 приведен график временного ряда `Mu Upload` – параметра расположения отправленного трафика. Очевидна ярко выраженная сезонность ряда, которая имеет место для всех четырнадцати временных рядов. Поскольку выборки анализируются с минимально возможным окном по времени, не исключено появление выбросов в ночные промежутки времени, например, как это произошло ночью с 20 на 21 день.

Стоит отметить, что ряды сильно скоррелированы. На рисунке 2.15 приведена корреляционная матрица всех рядов, для перечеркнутых корреляций гипотеза о равенстве нулю на уровне значимости $\alpha = 0.01$ не была отвергнута.

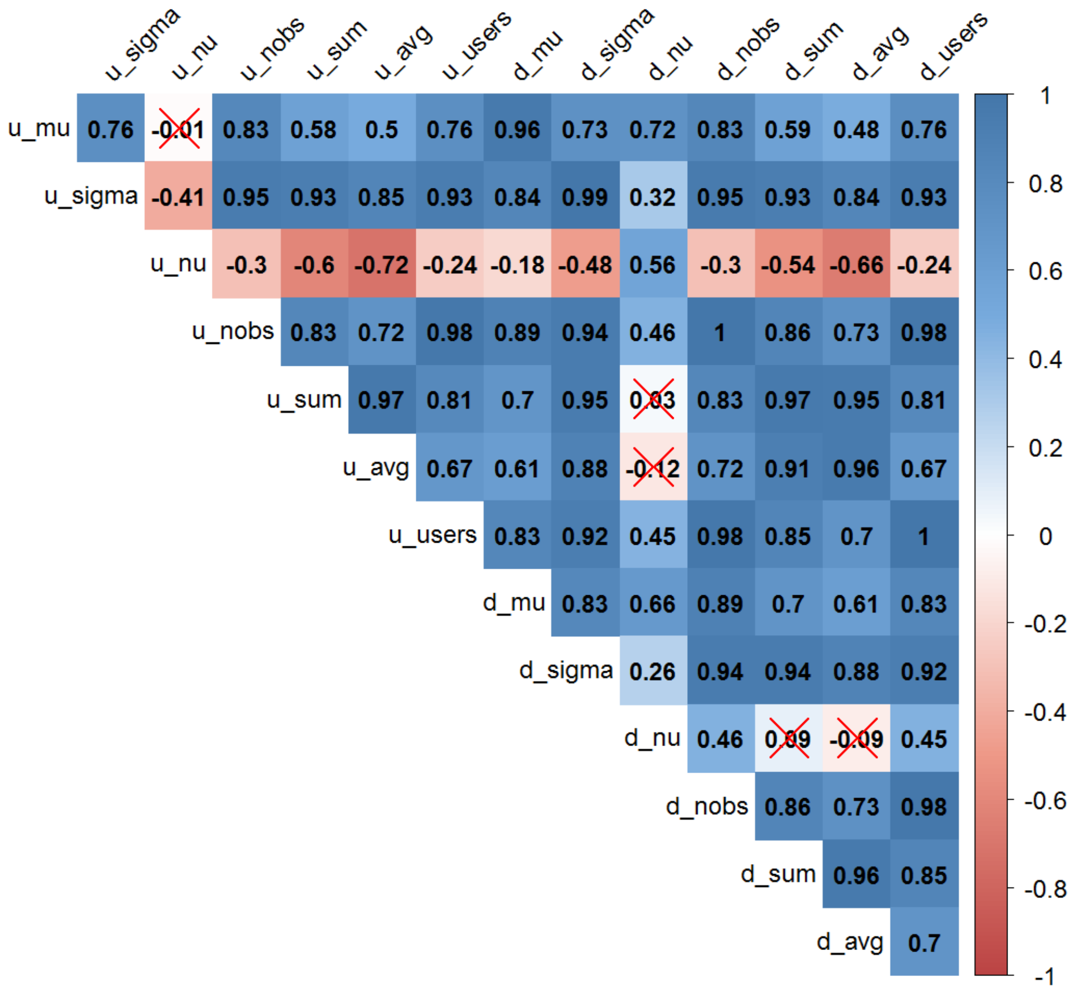



Рис. 2.15. Корреляционная матрица временных рядов

Для прогнозирования временных рядов были выбраны следующие три модели:

- Интегрированная модель авторегрессии – скользящего среднего с сезонностью – SARIMA$(p, d, q)(P, D, Q)[s]$
  Введем необходимые обозначения:

$B:\ Bx_t = x_{t-1}$ — оператор сдвига назад,
по индукции $B^k:\ B^k x_t = x_{t-k}$,
$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$,
$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q$,
$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \ldots - \Phi_P B^{Ps}$,
$\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \ldots + \Theta_Q B^{Qs}$,
$\Delta^d = (1-B)^d$ — оператор разности порядка $d$,
$\Delta_s^D = (1-B^s)^D$ — оператор сезонной разности порядка $D$,
$\varepsilon_t$ — процесс белого шума,
$c$ — некоторая константа.
Общий вид SARIMA$(p,d,q)(P,D,Q)[s]$ модели:

$$\phi(B)\Phi_P(B^s)\Delta^d\Delta_s^D x_t = c + \theta(B)\Theta_Q(B^s)\varepsilon_t.$$

Выбор параметров и оценивание коэффициентов SARIMA модели выполняется функцией `auto.arima` из библиотеки `forecast` на языке `R`.

- Регрессия на члены тригонометрического ряда Фурье до порядка $n$ с SARIMA$(p,d,q)(P,D,Q)[s]$ ошибками – Fourier$(n)$

$$x_t = c + \sum_{k=1}^{n}[a_k cos(kt) + b_k sin(kt)] + \eta_t,$$

где $\eta_t$ — SARIMA$(p,d,q)(P,D,Q)[s]$ ошибки,
$c$ — некоторая константа.

Выбор параметров и оценивание коэффициентов модели Фурье выполняется той же функцией `auto.arima` с добавлением регрессоров функцией `fourier` из библиотеки `forecast`.

- Векторная авторегрессия с сезонными фиктивными переменными – VAR$(p)$

$$X_t = C + \sum_{j=1}^{p}\Phi_j X_{t-j} + SD_t + \varepsilon_t,$$

где $X_t$ — векторный случайный процесс размерности $(n \times 1)$,
$\Phi_j: \ j = 1, \ldots, p$ — матрицы размера $(n \times n)$,
$D_t$ — центрированная сезонная фиктивная переменная — вектор размерности $((T-1) \times 1)$, такой что, если момент времени $t$ $i$-ый в сезоне $(1 \leqslant i \leqslant T-1)$, то на $i$-ой позиции вектора $D_t$ стоит $\frac{T-1}{T}$, а на остальных позициях $-\frac{1}{T}$, если момент времени $t$ последний в сезоне, то на всех позициях вектора $D_t$ стоит $-\frac{1}{T}$, где $T$ — длина сезона,
$S$ — матрица коэффициентов перед фиктивными переменными размера $(n \times (T-1))$,
$\varepsilon_t$ — многомерный $(n \times 1)$ процесс белого шума,
$C$ — некоторый вектор размерности $(n \times 1)$.

Оценивание коэффициентов модели VAR выполняется функцией `VAR` из библиотеки `vars` на языке `R`.

Предпосылки выбора моделей:

- SARIMA — классическое решение для моделирования временных рядов;
- регрессия на члены тригонометрического ряда Фурье была выбрана в силу очевидной сезонности данных;
- векторная авторегрессия была выбрана в силу скоррелированности временных рядов.

Правила выбора параметров моделей:

- параметры SARIMA модели выбираются по критерию Акаике [12];
- порядок членов ряда Фурье выбирается минимизацией RMSE на тренировочной выборке (оптимальный порядок для каждого ряда представлен в таблице 2.2), параметры SARIMA модели для остатков выбираются по критерию Акаике;

| | Mu | Sigma | Nu | Nobs | Sum | Avg | Users |
|---|---|---|---|---|---|---|---|
| Upload | 8 | 9 | 12 | 12 | 12 | 12 | 9 |
| Download | 12 | 12 | 10 | 12 | 12 | 11 | 9 |

Таблица 2.2. Оптимальный порядок модели Фурье

- порядок VAR модели выбирается с помощью функции `VARselect` из библиотеки `vars` на языке `R` голосованием по следующим информационным критериям [13]: критерий Акаике (AIC), критерий Ханнана-Куинна (HQ), критерий Шварца (SC) и критерий ошибки окончательного прогноза (FPE).

| AIC | HQ | SC | FPE |
|---|---|---|---|
| 2 | 2 | 1 | 2 |

Таблица 2.3. Выбор параметра VAR

Согласно таблице 2.3, имеет смысл оценить модели VAR($p$) для $p = 1, 2$, однако также будет проверена и модель более высокого порядка ($p = 3$). Таким образом, для каждого ряда обучается пять моделей: SARIMA, Fourier($n$), VAR(1), VAR(2), VAR(3). На рисунках 2.16 – 2.20 приведены результаты прогнозирования ряда `Mu Upload` всеми указанными моделями. Оранжевым цветом обозначен 95% доверительный интервал для прогноза.

Наконец, вычислим ошибку прогноза на тестовой выборке по нормированным данным. Для этого сначала нормируем тренировочную выборку по формуле $\mathrm{norm}(x) = \frac{x - \min(\mathrm{train})}{\max(\mathrm{train}) - \min(\mathrm{train})}$, получим прогноз по нормированной тренировочной выборке, нормируем тестовую выборку по той же формуле, что и тренировочную, и вычислим метрику RMSE по формуле (2.3) между прогнозом и нормированной тестовой выборкой. В таблице 2.4 приведены описанные вычисления для каждого временного ряда и каждой модели.

$$\mathrm{RMSE}(\vec{x}, \vec{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}. \tag{2.3}$$

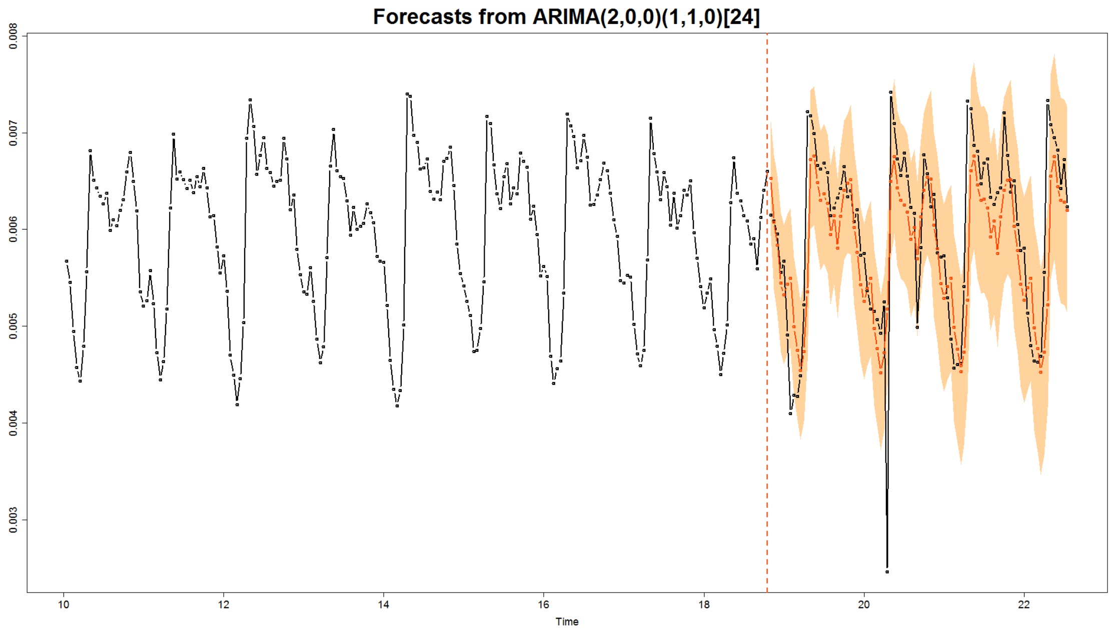



Рис. 2.16. Прогнозирование ряда `Mu Upload` моделью SARIMA

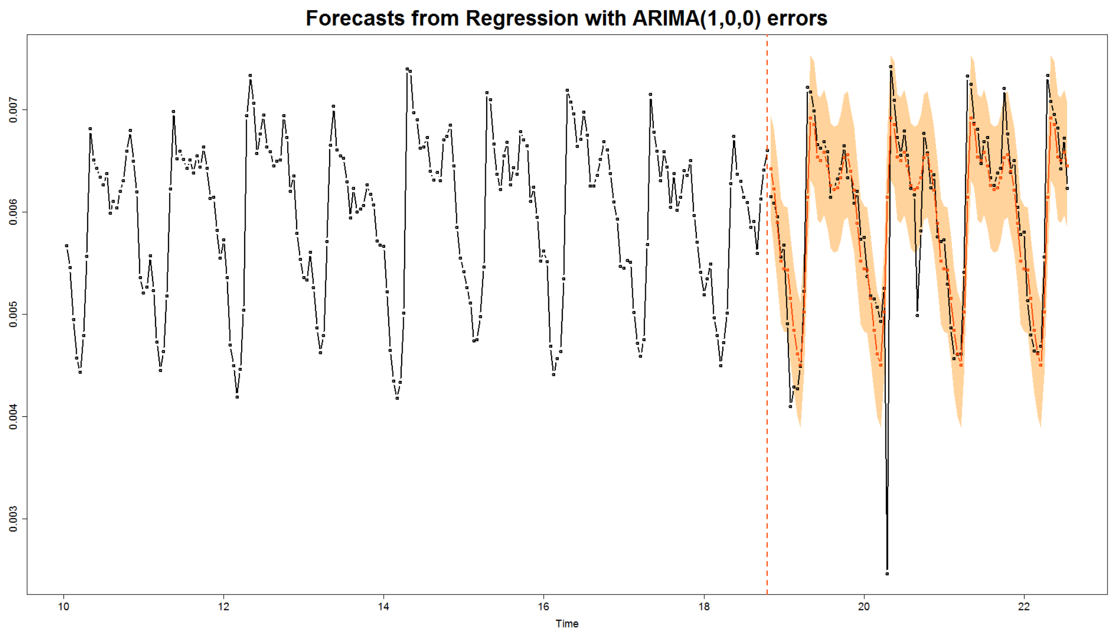



Рис. 2.17. Прогнозирование ряда `Mu Upload` моделью Fourier

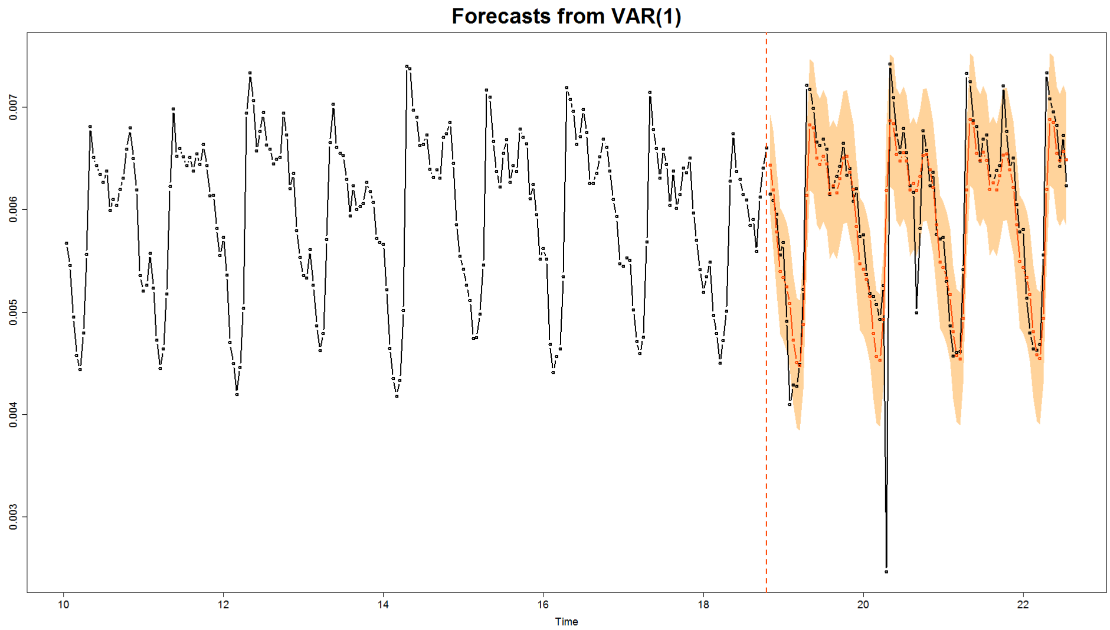



Рис. 2.18. Прогнозирование ряда `Mu Upload` моделью VAR(1)

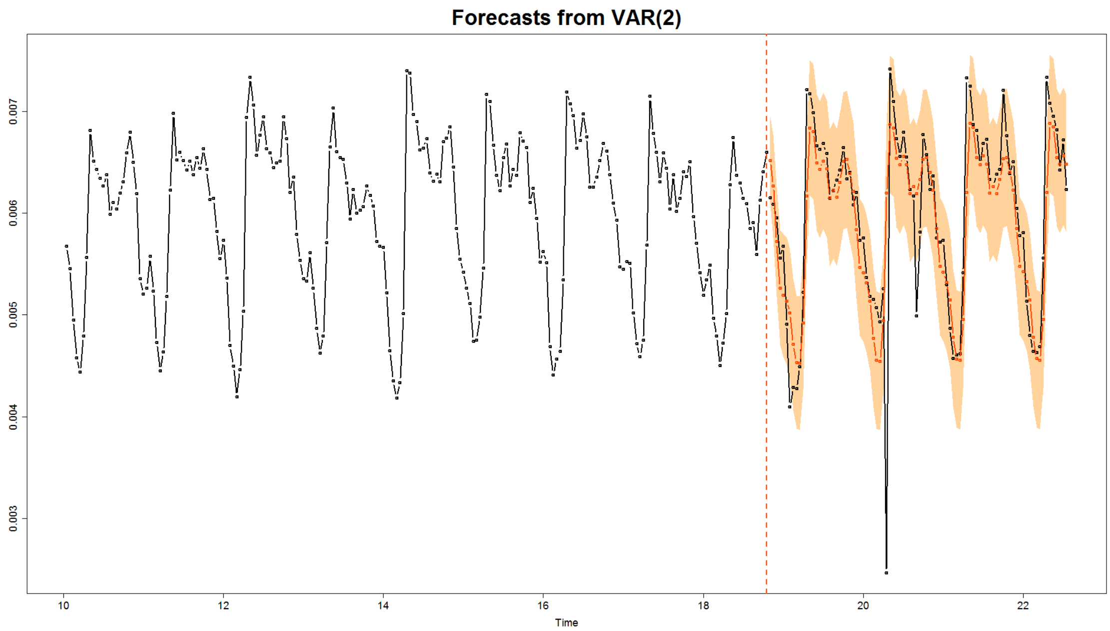



Рис. 2.19. Прогнозирование ряда `Mu Upload` моделью VAR(2)

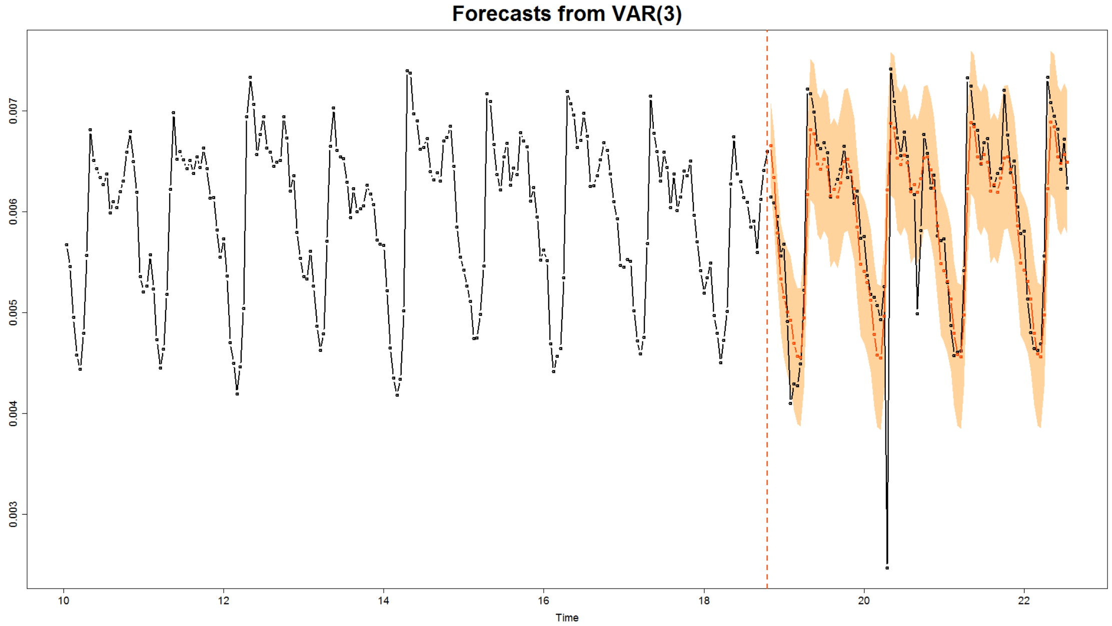



Рис. 2.20. Прогнозирование ряда `Mu Upload` моделью VAR(3)

| Тип | Ряд | SARIMA | Fourier | VAR1 | VAR2 | VAR3 |
|---|---|---|---|---|---|---|
| Upload | Mu | 0.192 | **0.1645** | 0.1652 | 0.1649 | 0.1646 |
| | Sigma | 0.1505 | 0.0896 | 0.0897 | 0.0864 | **0.0846** |
| | Nu | 0.1467 | **0.1423** | 0.1488 | 0.1472 | 0.1467 |
| | Nobs | 0.1767 | 0.1281 | 0.1232 | 0.1205 | **0.12** |
| | Sum | 0.1175 | **0.1118** | 0.1256 | 0.124 | 0.1235 |
| | Avg | 0.1196 | **0.1156** | 0.1368 | 0.1346 | 0.134 |
| | Users | 0.1797 | 0.1302 | 0.1266 | 0.1234 | **0.1226** |
| Download | Mu | 0.2284 | 0.1683 | 0.1688 | 0.1676 | **0.1675** |
| | Sigma | 0.1198 | 0.085 | 0.0818 | 0.0789 | **0.0776** |
| | Nu | 0.1974 | **0.1632** | 0.1743 | 0.173 | 0.173 |
| | Nobs | 0.1825 | 0.1292 | 0.1245 | 0.1217 | **0.1212** |
| | Sum | 0.1132 | 0.1124 | 0.1126 | 0.111 | **0.11** |
| | Avg | 0.1045 | **0.0993** | 0.1147 | 0.1136 | 0.1126 |
| | Users | 0.21 | 0.1436 | 0.1285 | 0.1252 | **0.1244** |

Таблица 2.4. Ошибка на нормированных данных с одночасовым окном

Для шести из четырнадцати анализируемых рядов лучшей в смысле минимальной ошибки оказалась модель Fourier, для остальных восьми — VAR(3). Выбрать лучшую из этих двух моделей по таблице 2.4 затруднительно: на ряде `Avg Upload` модель Fourier выиграла у VAR(3) около 2% в точности, на ряде `Users Download` ситуация противоположная. В среднем VAR(3) оказывается точнее всего лишь на 0.006%. Однако не стоит забывать, что в реальных задачах ключевую роль играет не только точность прогнозирования, но и время, за которое строится прогноз. В таблице 2.5 приведено время в секундах, затраченное каждой моделью на обучение и получение прогноза. Модель векторной авторегресии отрабатывает в сотни раз быстрее и может быть использована при анализе в режиме реального времени.

<table>
<tr><th>Тип</th><th>Ряд</th><th>SARIMA</th><th>Fourier</th><th>VAR1</th><th>VAR2</th><th>VAR3</th></tr>
<tr><td rowspan="7">Upload</td><td>Mu</td><td>25.49</td><td>17.17</td><td rowspan="14">0.22</td><td rowspan="14">0.24</td><td rowspan="14">0.28</td></tr>
<tr><td>Sigma</td><td>62.96</td><td>376.89</td></tr>
<tr><td>Nu</td><td>36.44</td><td>48.61</td></tr>
<tr><td>Nobs</td><td>87.86</td><td>69.78</td></tr>
<tr><td>Sum</td><td>44.85</td><td>112.39</td></tr>
<tr><td>Avg</td><td>78.19</td><td>699.75</td></tr>
<tr><td>Users</td><td>133.02</td><td>273.52</td></tr>
<tr><td rowspan="7">Download</td><td>Mu</td><td>56.75</td><td>56.21</td></tr>
<tr><td>Sigma</td><td>129.44</td><td>64.73</td></tr>
<tr><td>Nu</td><td>36.57</td><td>197.05</td></tr>
<tr><td>Nobs</td><td>78.68</td><td>64.37</td></tr>
<tr><td>Sum</td><td>55.4</td><td>605.13</td></tr>
<tr><td>Avg</td><td>58.72</td><td>1039.48</td></tr>
<tr><td>Users</td><td>135.9</td><td>1336.98</td></tr>
</table>

Таблица 2.5. Время обучения и прогнозирования с одночасовым окном

Ранее уже было отмечено, что вышеописанный анализ может быть повторён с любыми другими возможными агрегациями трафика по времени. На рисунке 2.21 показан результат прогнозирования моделью Fourier ряда `Mu Upload` с четырёхчасовом окном. По мере увеличения окна по времени выбросы в ночные периоды времени пропадают, но вместе с тем исчезают дневные паттерны и значительно сокращается длина временного ряда ряда.

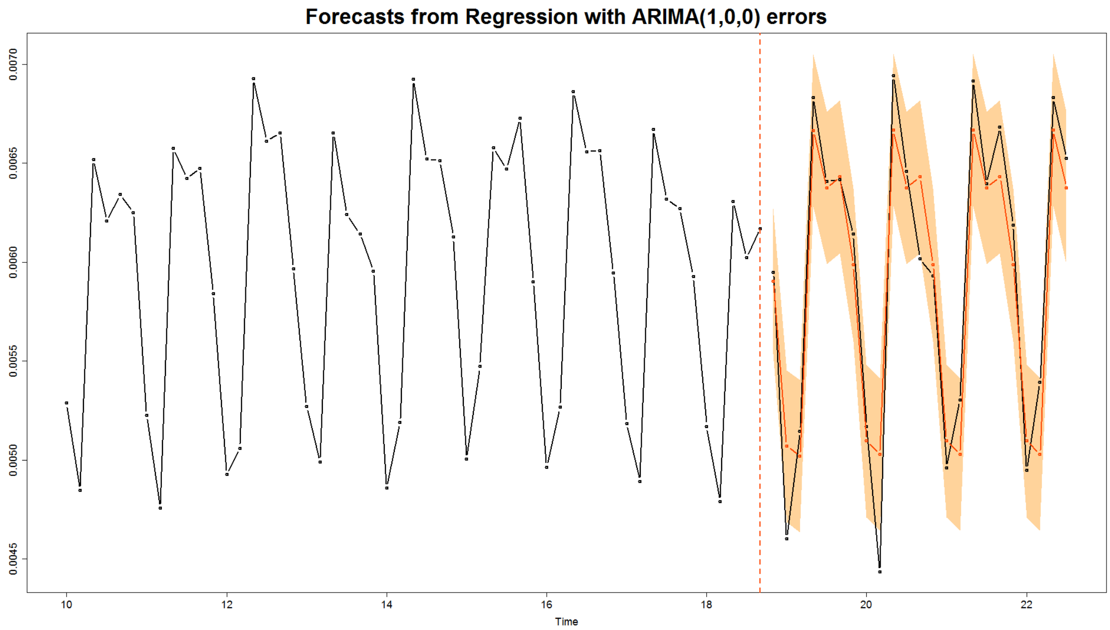



Рис. 2.21. Прогнозирование ряда `Mu Upload` моделью Fourier с четырёхчасовым окном

В таблице 2.6 приведены ошибки прогнозирования с четырёхчасовым окном. Модель SARIMA на некоторых рядах оказывается лучшей, однако её проигрыши в точности на других рядах слишком велики. Это приводит к тому, что модель SARIMA в среднем прогнозирует на 4.71% хуже модели Fourier и на 4.46% хуже модели VAR(1). Модели Fourier и VAR(1) в среднем прогнозируют практически одинаково — Fourier выигрывает в точности 0.2%.

Поскольку с увеличением окна по времени существенно сократилась длина временного ряда и его период, модели SARIMA и Fourier стали прогнозировать гораздо быстрее и теперь уступают по быстродействию модели векторной авторегрессии не в сотни, а в десятки раз. Информация о времени в секундах, затраченном на обучение и прогнозирование с четырёхчасовым окном, содержится в таблице 2.7.

| Тип | Ряд | SARIMA | Fourier | VAR1 | VAR2 |
|---|---|---|---|---|---|
| Upload | Mu | 0.1476 | **0.1124** | 0.1133 | 0.3001 |
| | Sigma | 0.1771 | 0.0844 | **0.0802** | 0.2694 |
| | Nu | 0.1488 | **0.1167** | 0.1209 | 0.4055 |
| | Nobs | 0.2649 | **0.2254** | 0.2276 | 0.2849 |
| | Sum | **0.1531** | 0.1561 | 0.1642 | 0.1726 |
| | Avg | **0.1593** | 0.1603 | 0.1755 | 0.1906 |
| | Users | 0.2126 | **0.1164** | 0.1168 | 0.231 |
| Download | Mu | 0.1888 | 0.1081 | **0.1051** | 0.289 |
| | Sigma | 0.139 | 0.0786 | **0.0758** | 0.1788 |
| | Nu | 0.2499 | **0.1589** | 0.1629 | 0.4238 |
| | Nobs | 0.2672 | **0.2254** | 0.2279 | 0.2844 |
| | Sum | **0.159** | 0.1617 | 0.1624 | 0.1667 |
| | Avg | **0.1539** | 0.1589 | 0.1657 | 0.1946 |
| | Users | 0.218 | **0.1156** | 0.1163 | 0.2463 |

Таблица 2.6. Ошибка на нормированных данных с четырёхчасовым окном

| Тип | Ряд | SARIMA | Fourier | VAR1 | VAR2 |
|---|---|---|---|---|---|
| Upload | Mu | 0.93 | 0.63 | 0.07 | 0.1 |
| | Sigma | 1.52 | 0.9 | | |
| | Nu | 0.66 | 0.75 | | |
| | Nobs | 1.36 | 0.55 | | |
| | Sum | 0.8 | 0.65 | | |
| | Avg | 4.5 | 0.62 | | |
| | Users | 3.19 | 0.78 | | |
| Download | Mu | 0.3 | 1.06 | | |
| | Sigma | 1.9 | 1.55 | | |
| | Nu | 0.68 | 0.46 | | |
| | Nobs | 1.53 | 0.69 | | |
| | Sum | 1.59 | 0.71 | | |
| | Avg | 0.8 | 0.81 | | |
| | Users | 2.87 | 0.8 | | |

Таблица 2.7. Время обучения и прогнозирования с четырёхчасовым окном

## 2.5 Выявление аномалий

В этом разделе речь пойдет о статистической процедуре выявления аномальных наблюдений в выборках из обобщённого гамма-распределения, описанной в статье [14]. Всюду в этом разделе будет использована параметризация плотности обобщённого гамма-распределения, задаваемая формулой (2.2).

**Предложение 2.** *Пусть $V_1, \ldots, V_m$ независимая выборка из обобщённого гамма-распределения с некоторыми параметрами $r > 0,\ \gamma > 0,\ \mu > 0$, $V_1 \geqslant V_j,\ \forall j \geqslant 2$,*

$$\mathcal{R} = \frac{(m-1)V_1^\gamma}{V_2^\gamma + \cdots + V_m^\gamma}. \tag{2.4}$$

*Тогда при условии, что верна гипотеза $H_0$: «значение $V_1$ не является аномально большим», статистика $\mathcal{R}$ имеет распределение Снедекора-Фишера с параметрами $r$ и $(m-1)r$.*

В упомянутой статье в силу специфики предметной области анализируются выборки, имеющие обобщённое гамма-распределение с положительным параметром $\gamma$. В задачах анализа мобильного трафика встречаются случаи как положительного, так и отрицательного параметра $\gamma$.

Для начала продемонстрируем работу предложенного статистического теста на выборке объемов трафика с положительным параметром $\gamma$. Зафиксируем уровень значимости $\alpha = 0.05$ и вычислим по формуле (2.4) значения статистики $R_i$ для каждого $i$-ого наблюдения в выборке. Те наблюдения, для которых $\mathbb{P}(\mathcal{R} \geqslant R_i) < \alpha$, будут признаны аномальными. На рисунках 2.22 – 2.24 представлены результаты выявления аномальных наблюдений в выборках с положительным параметром $\gamma$. Аномальные наблюдения помечены красным.

Поскольку выборки объёмов трафика зачастую имеют обобщённое гамма-распределение с отрицательным параметром $\gamma$, необходимо обобщить вышеописанный статистический тест на случай пара-

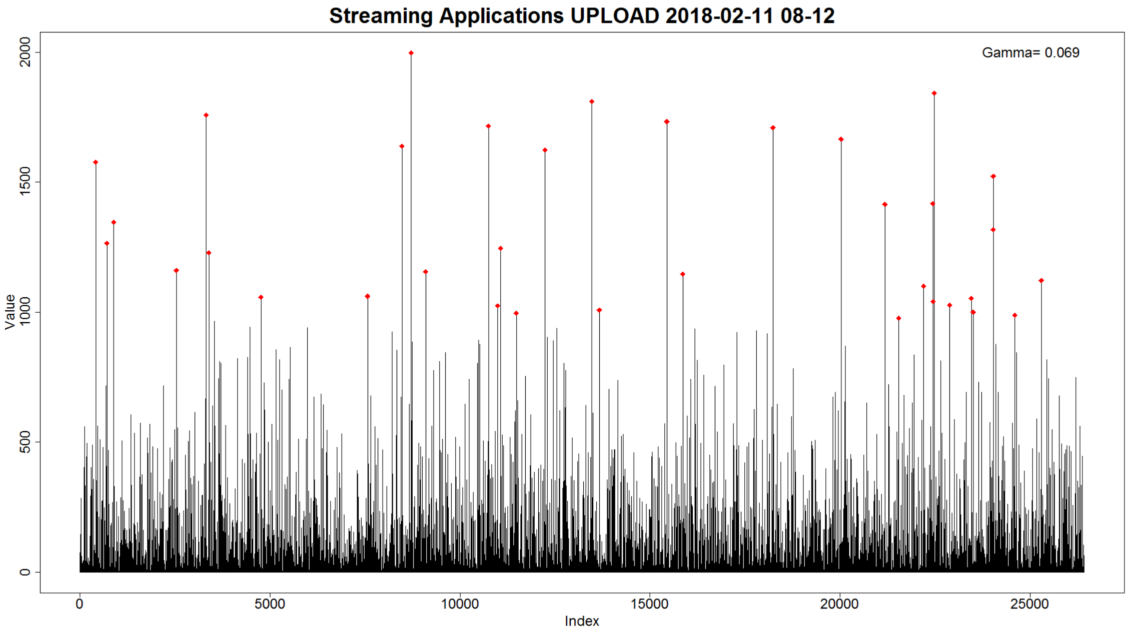



Рис. 2.22. Выявление аномальных объёмов отправленного трафика по приложению `Streaming Applications` за четыре часа одного дня

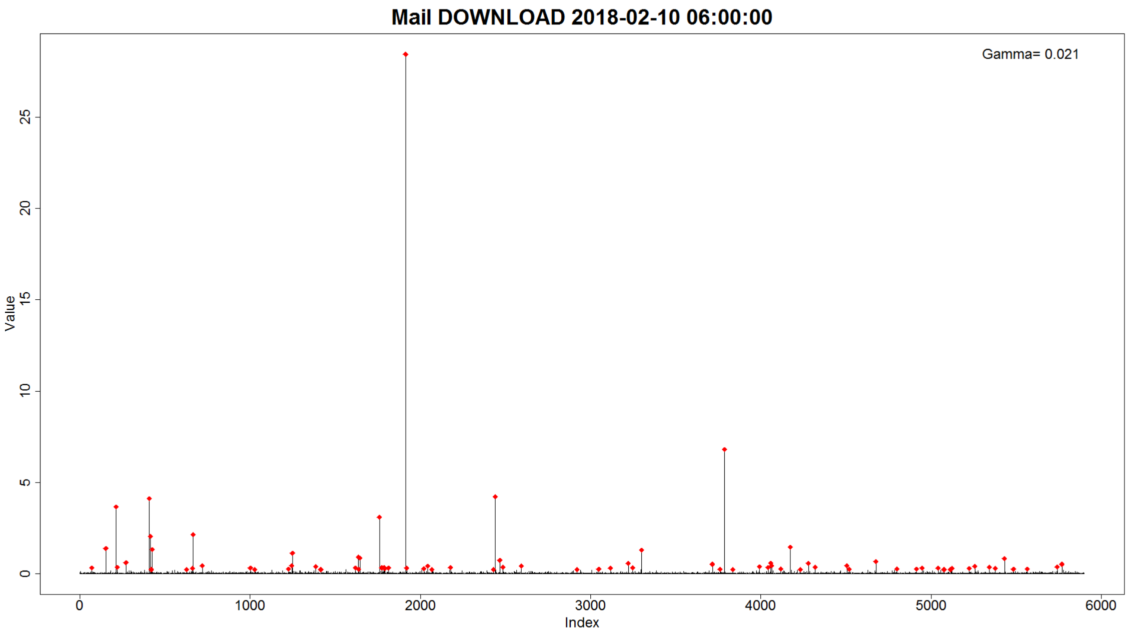



Рис. 2.23. Выявление аномальных объёмов полученного трафика по приложению `Mail` за один час одного дня

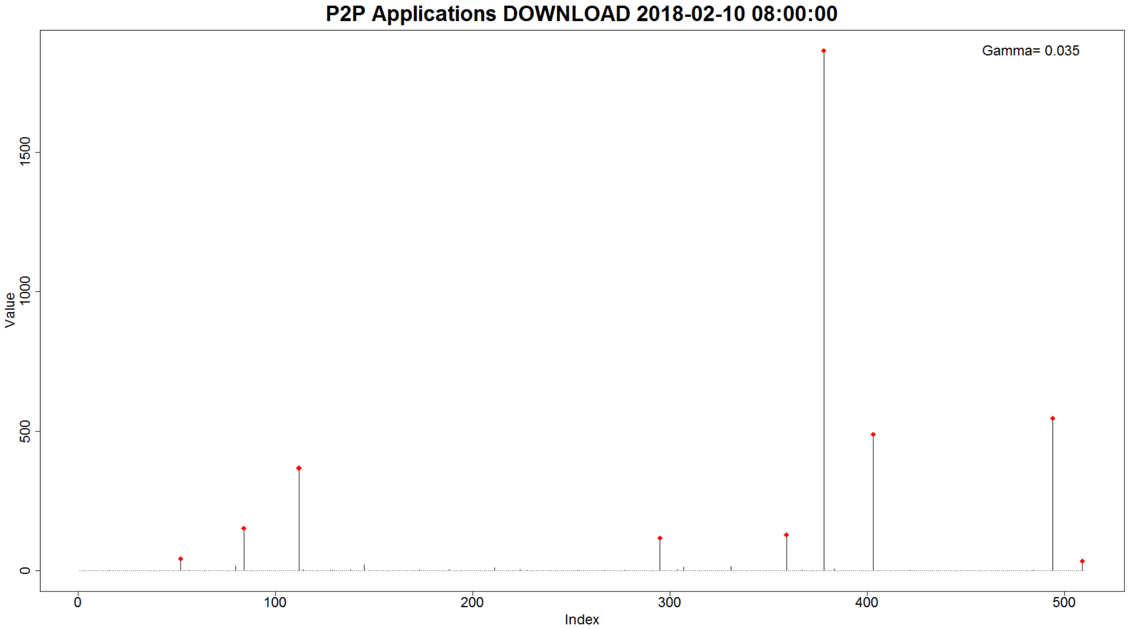



Рис. 2.24. Выявление аномальных объёмов полученного трафика по приложению `P2P Applications` за один час одного дня

метра $\gamma$ произвольного знака. Заметим, что независимо от знака параметра $\gamma$ статистика $\mathcal{R}$ будет иметь распределение Снедекора-Фишера с параметрами $r$ и $(m-1)r$ (см. доказательство, приведенное в статье [14]).

Рассмотрим случай $\gamma < 0$. В силу убывания функции $f(x) = x^\gamma$ на положительной полупрямой и аналитического вида статистики $\mathcal{R}$ бóльшим значениям наблюдений в выборке будут соответствовать меньшие значения статистики. Из этого рассуждения возникает две идеи модификации статистического теста для случая $\gamma < 0$:

- Рассматривать прежнюю статистику $\mathcal{R}$, для каждого $i$-ого наблюдения вычислять значение статистики $R_i$ по формуле (2.4) и признавать аномально большими те наблюдения, для которых $\mathbb{P}(\mathcal{R} < R_i) < \alpha$;

- Рассматривать статистику $\overline{\mathcal{R}} = \frac{1}{\mathcal{R}}$ для каждого $i$-ого наблюдения вычислять значение статистики $\overline{R_i} = \frac{1}{R_i}$ и признавать аномально большими те наблюдения, для которых $\mathbb{P}(\overline{\mathcal{R}} \geqslant \overline{R_i}) < \alpha$.

Нетрудно убедиться в том, что два этих подхода эквивалентны:

$$\mathbb{P}(\overline{\mathcal{R}} \geqslant \overline{R_i}) < \alpha \Leftrightarrow \mathbb{P}(\frac{1}{\mathcal{R}} \geqslant \frac{1}{R_i}) < \alpha \Leftrightarrow \mathbb{P}(\mathcal{R} \leqslant R_i) < \alpha$$

Остановимся на втором варианте для единообразия выбора критической области «на бесконечности». Тогда обобщённая статистика будет иметь вид

$$\hat{\mathcal{R}} = \left(\frac{(m-1)V_1^{\gamma}}{V_2^{\gamma} + \cdots + V_m^{\gamma}}\right)^{\mathrm{sgn}(\gamma)}. \tag{2.5}$$

**Предложение 3.** *Пусть случайная величина $X$ имеет распределение Снедекора-Фишера с параметрами $d_1$ и $d_2$. Тогда случайная величина $\frac{1}{X}$ имеет распределение Снедекора-Фишера с параметрами $d_2$ и $d_1$.*

В силу выражения (2.5), предложения 2 и предложения 3 заключаем:

$$\hat{\mathcal{R}} \sim F(r,\ (m-1)r) \text{ при } \gamma > 0,$$

$$\hat{\mathcal{R}} \sim F((m-1)r,\ r) \text{ при } \gamma < 0.$$

На рисунках 2.25 – 2.28 представлены результаты выявления аномальных наблюдений в выборках с отрицательным параметром $\gamma$.

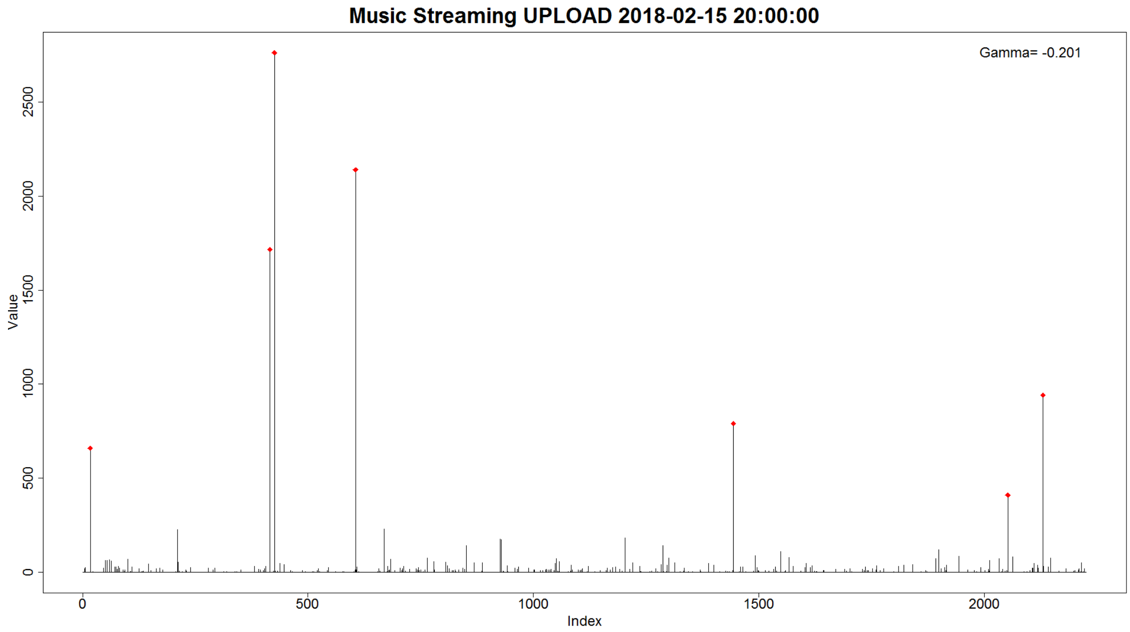



Рис. 2.25. Выявление аномальных объёмов отправленного трафика по приложению `Music Streaming` за один час одного дня

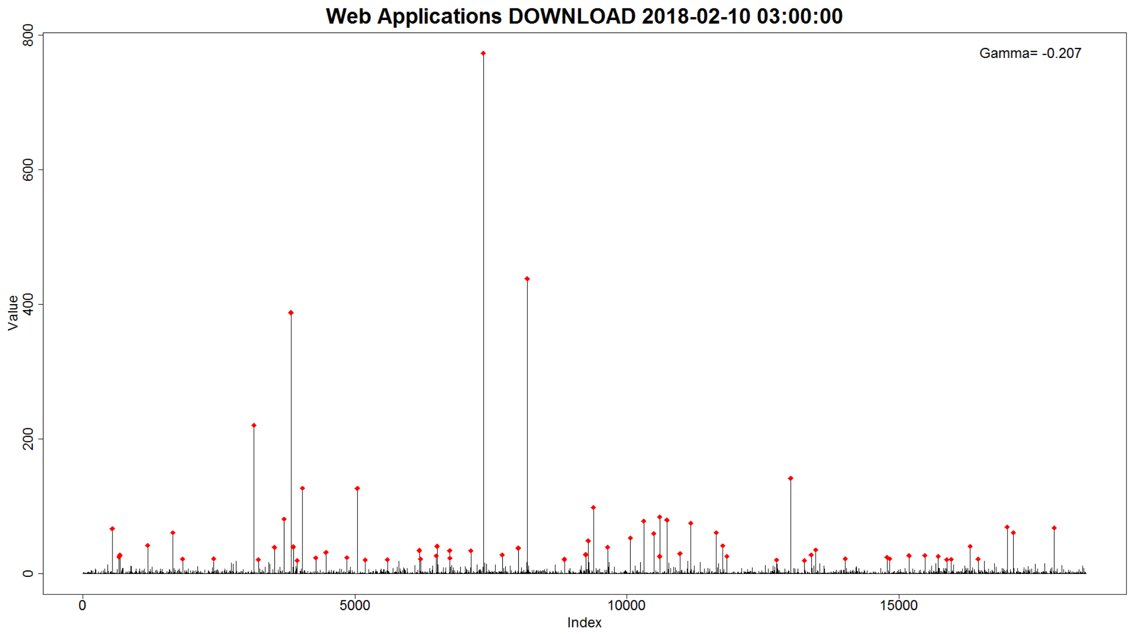



Рис. 2.26. Выявление аномальных объёмов полученного трафика по приложению `Web Applications` за один час одного дня

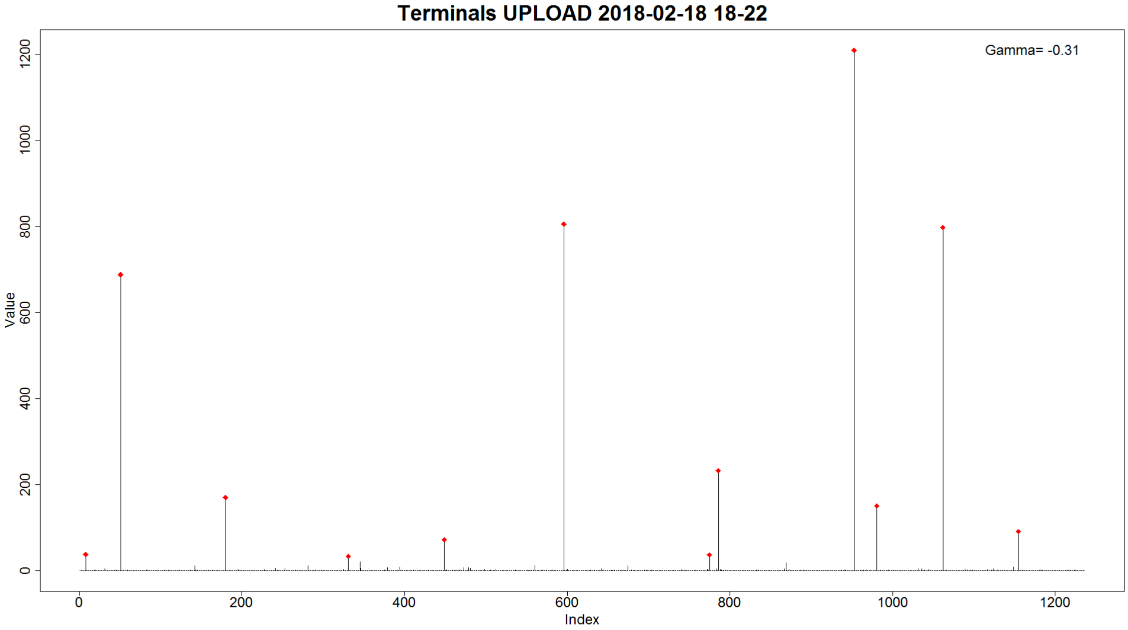



Рис. 2.27. Выявление аномальных объёмов отправленного трафика по приложению `Terminals` за четыре часа одного дня

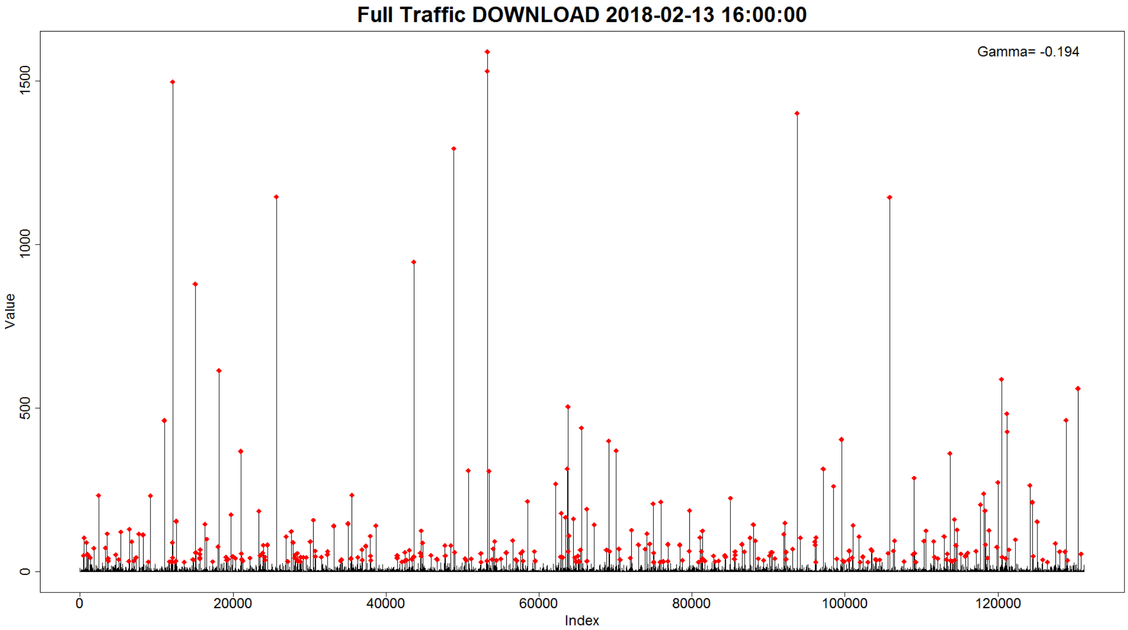



Рис. 2.28. Выявление аномальных объёмов полученного трафика без разбиения на приложения за один час одного дня

# Глава 3

# Анализ характеристик загруженности соты

Эта глава посвящена статистическому анализу характеристик загруженности соты, а именно анализу суммарного трафика, количества уникальных пользователей и среднего трафика, приходящегося на уникального пользователя.

## 3.1 Прогнозирование суммарного и среднего трафика

Техники прогнозирования временных рядов, получающихся при различных агрегациях трафика по времени, были описаны в разделе 2.4. В этом разделе, как и ранее, будет рассматриваться агрегация трафика с минимальным окном по времени (одночасовым). На рисунках 3.1 – 3.4 приведены результаты прогнозирования непосредственно объемов суммарного и среднего трафика моделью с минимальной ошибкой на тестовой выборке. Ошибка прогнозирования приведена в таблице 2.4, порядок модели регрессии на члены ряда Фурье – в таблице 2.2. Для оценки загруженности соты можно использовать точечные прогнозы, однако более надежный прогноз даст верхняя граница 95% доверительного интервала.

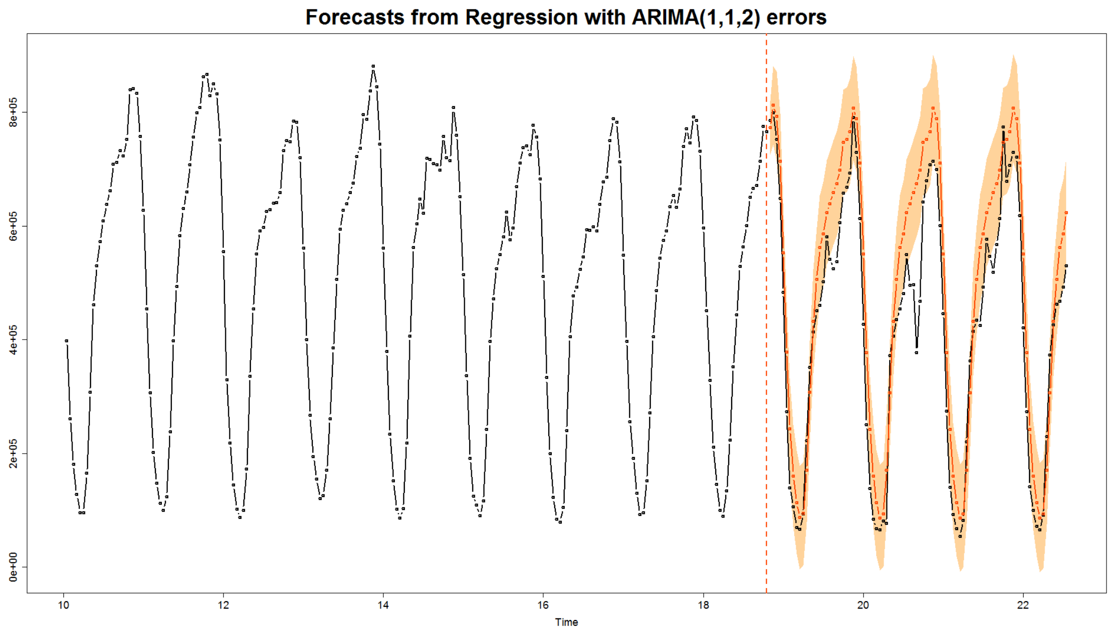



Рис. 3.1. Прогнозирование суммарного отправленного трафика

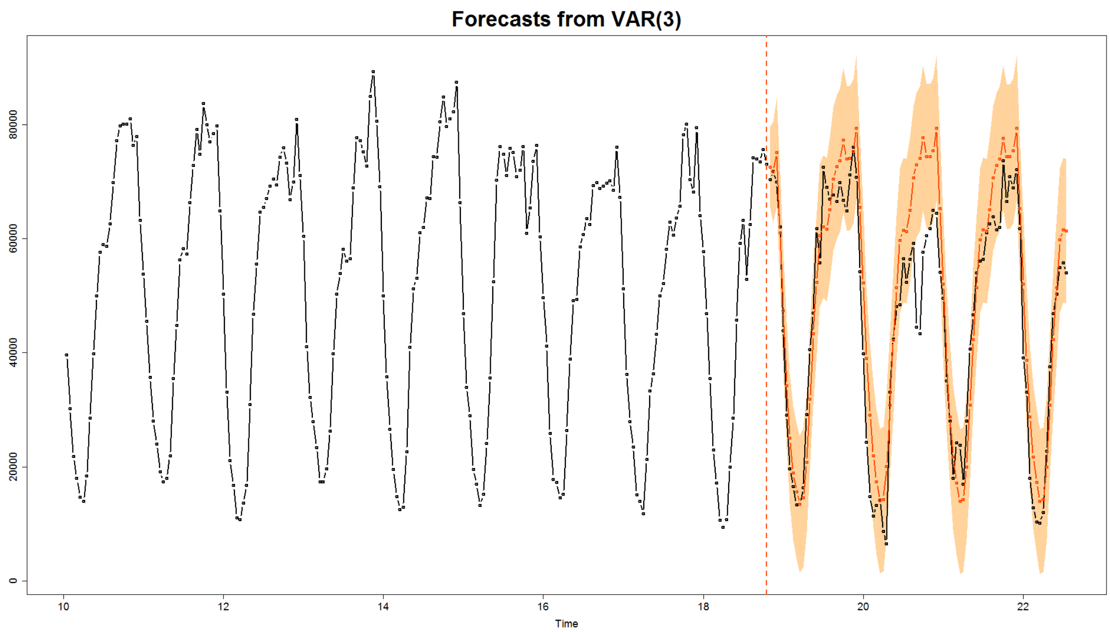



Рис. 3.2. Прогнозирование суммарного полученного трафика

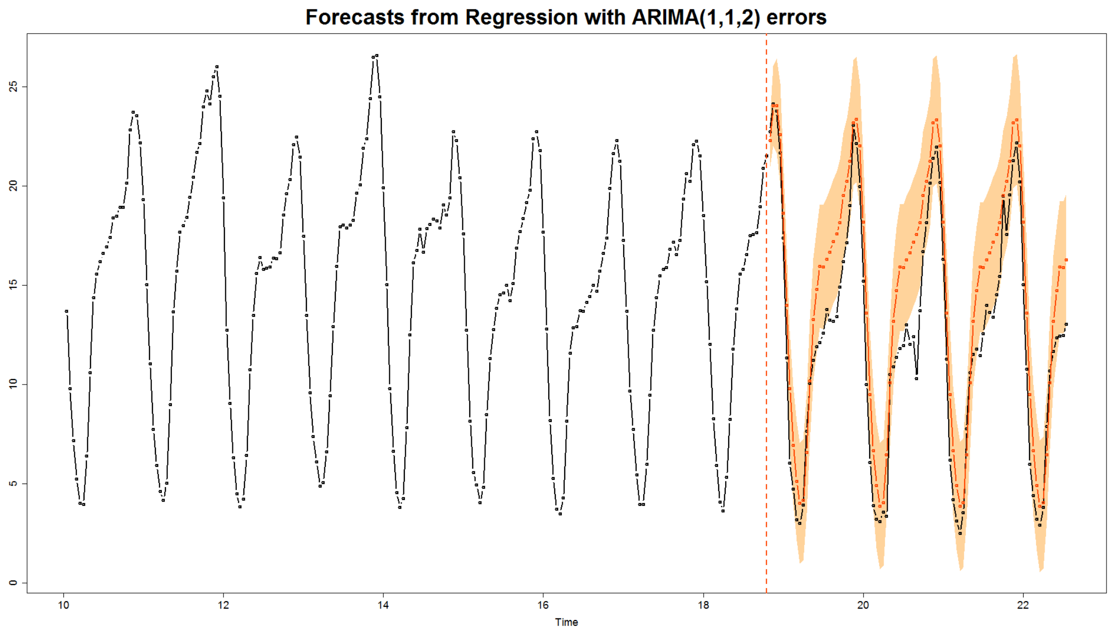



Рис. 3.3. Прогнозирование среднего отправленного трафика

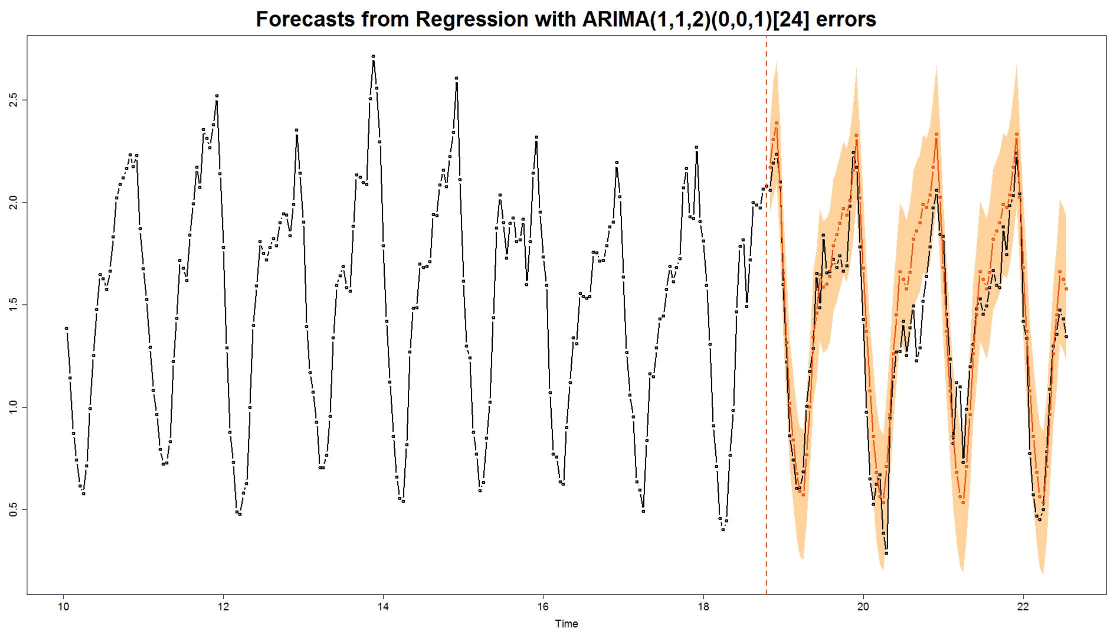



Рис. 3.4. Прогнозирование среднего полученного трафика

## 3.2 Анализ количества уникальных пользователей

Количество уникальных пользователей может быть спрогнозировано подобно тому, как это было сделано в разделе 2.4 для параметров обобщённого гамма-распределения и в разделе 3.1 для объёмов суммарного и среднего трафика. Результаты прогнозирования количества уникальных пользователей по отправленному и полученному трафику лучшей в смысле минимальной ошибки на тестовой выборке моделью приведены на рисунках 3.5 – 3.6.

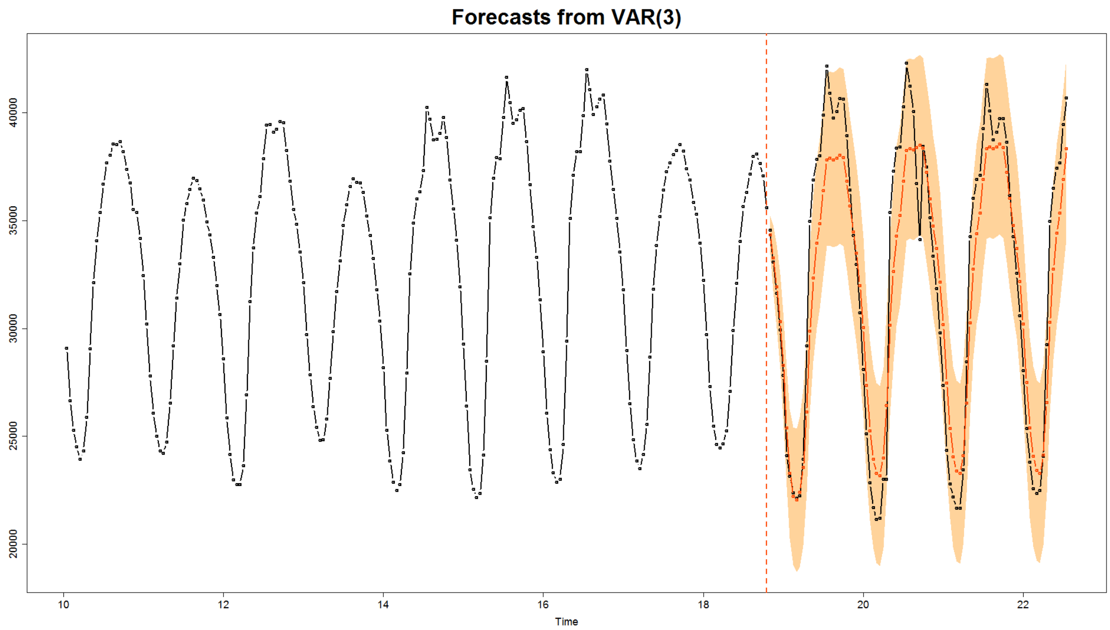



Рис. 3.5. Прогнозирование количества уникальных пользователей по отправленному трафику

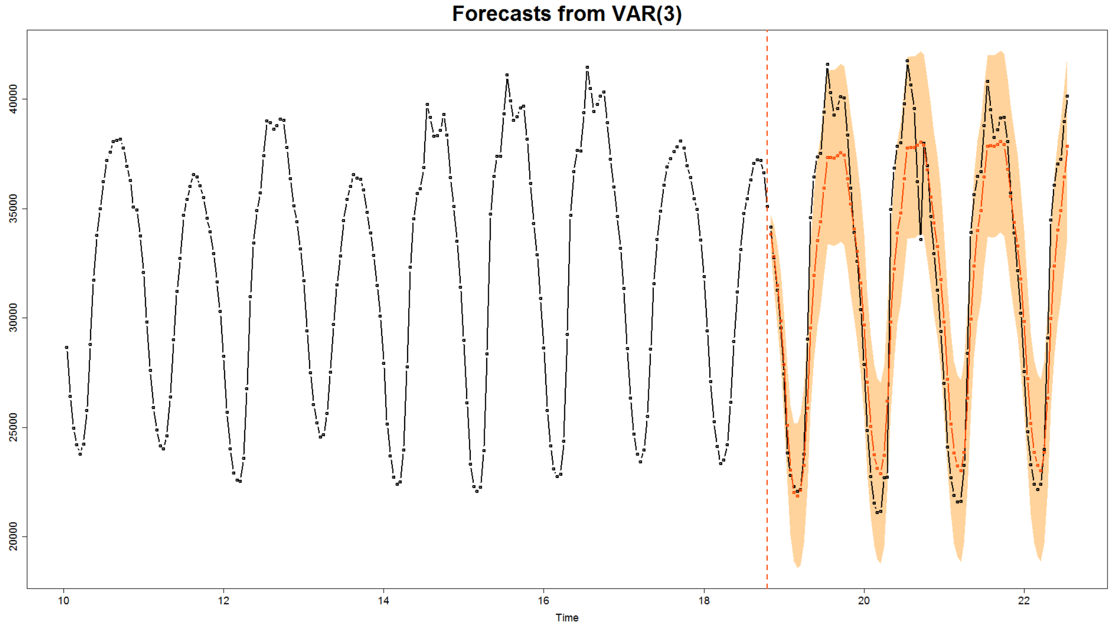



Рис. 3.6. Прогнозирование количества уникальных пользователей по полученному трафику

Поскольку данные мобильного оператора собирались за два временных промежутка (10.02.2018 – 22.02.2018 и 01.03.2018 – 04.03.2018), отдельный интерес представляет проверка однородности распределений количества уникальных пользователей за первый и второй интервалы времени. На рисунке 3.7 построен временной ряд количества уникальных пользователей по отправленному трафику до и после перерыва (перерыв обозначен вертикальной прерывистой линией). Для проверки гипотезы однородности сформируем выборки количества уникальных пользователей за первый и второй промежуток времени и проведем тест Колмогорова-Смирнова. Гистограммы распределений для отправленного трафика представлены на рисунке 3.8, p-значение теста на однородность и графики эмпирических функций распределения – на рисунке 3.9. Ананлогичный график эмпирических функций распределения для полученного трафика приведен на рисунке 3.10.

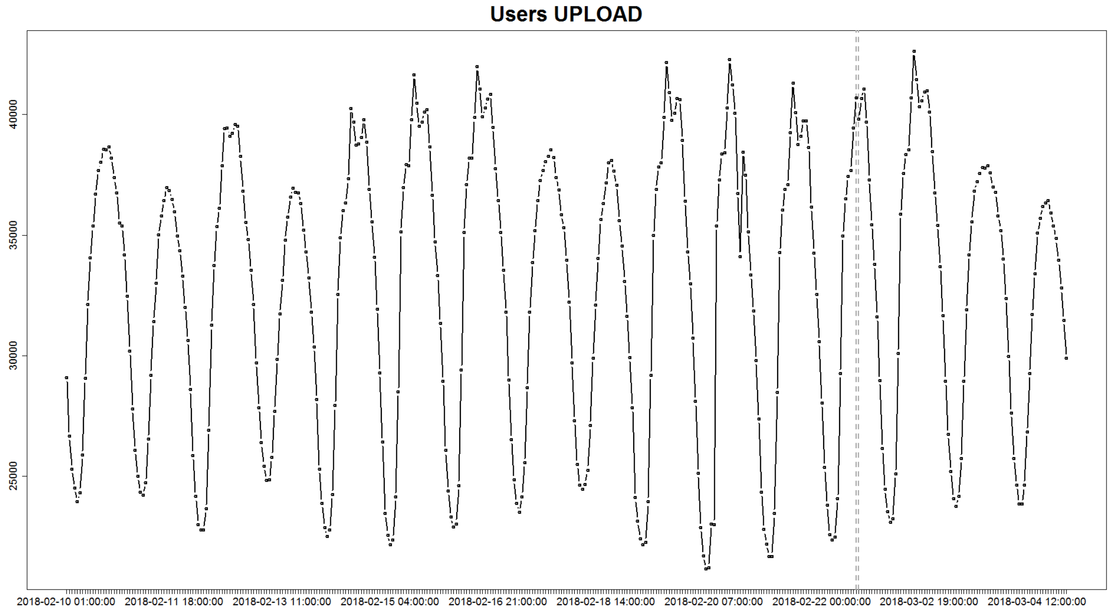



Рис. 3.7. Количество уникальных пользователей за час по отправленному трафику

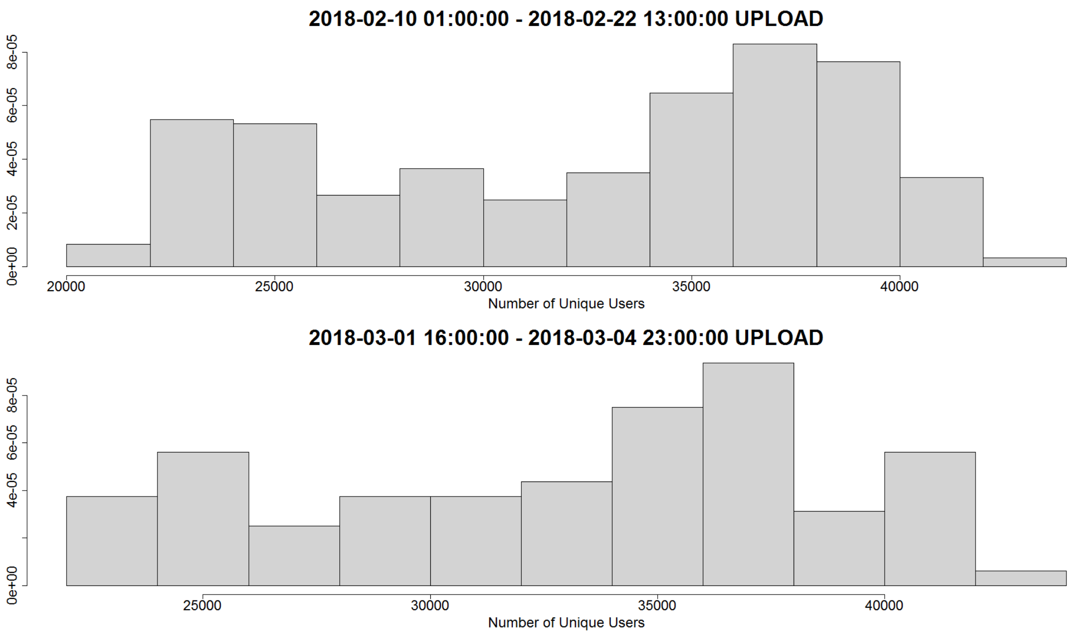



Рис. 3.8. Распределения количества уникальных пользователей до и после перерыва

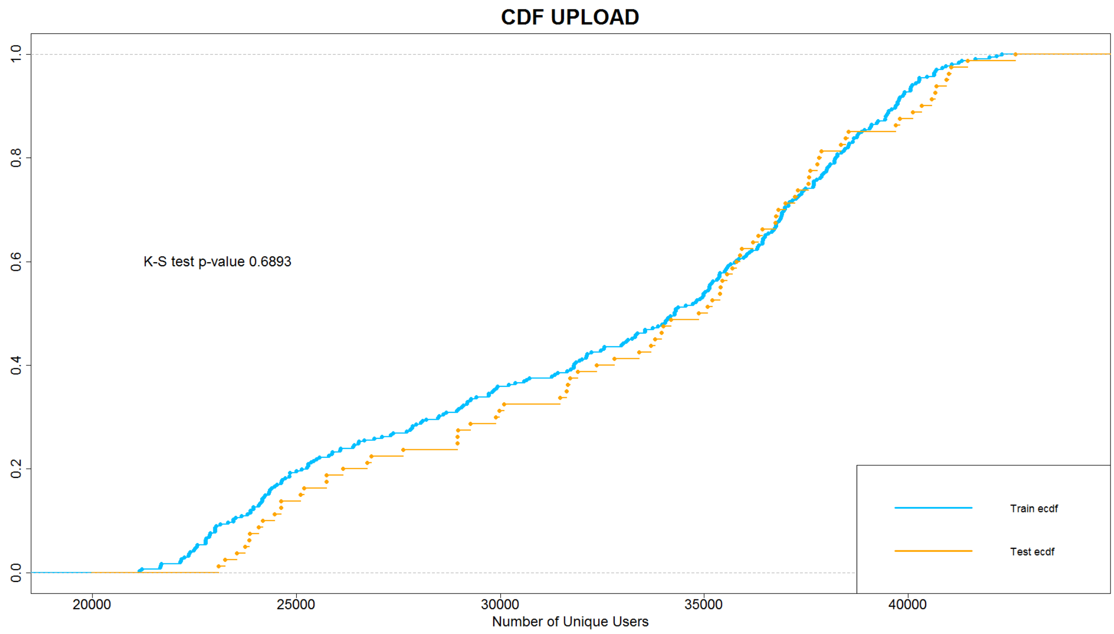



Рис. 3.9. Эмпирические функции распределения количества пользователей по отправленному трафику

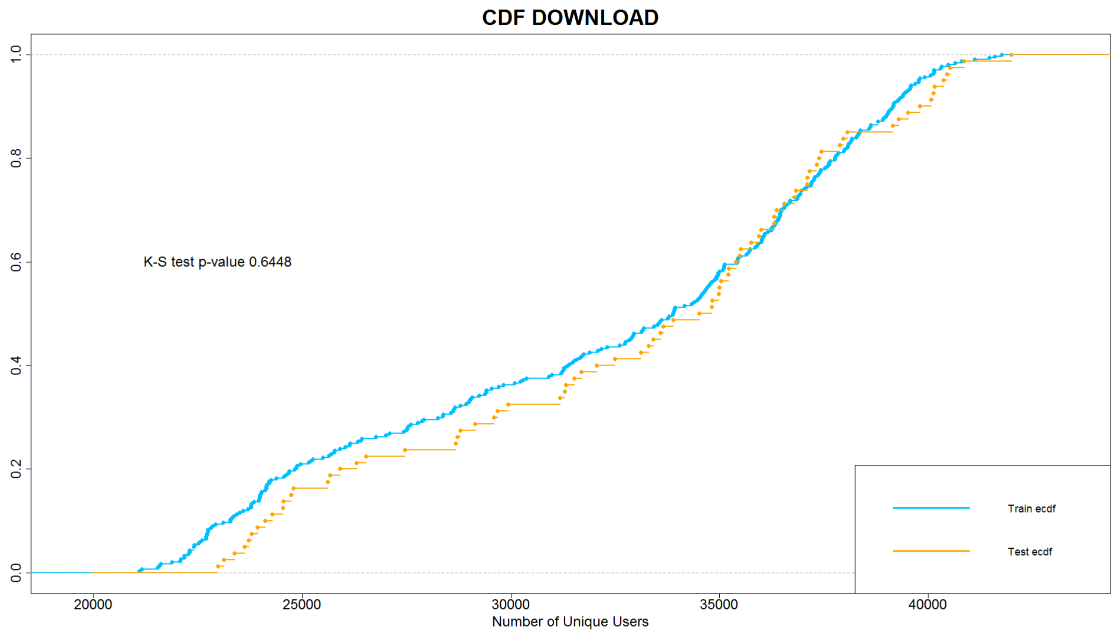



Рис. 3.10. Эмпирические функции распределения количества пользователей по полученному трафику

На гистограммах ярко выражены два пика, соответствующие ночному и дневному периодам времени, поэтому при необходимости аппроксимации этих распределений некоторым семейством вероятностных распределений следует выбирать двухкомпонентные смеси распределений. Поскольку p-значение тестов на однородность больше любого адекватного уровня значимости ($\alpha = 0.05;\ 0.01$), гипотеза об однородности не отвергается.

## 3.3 Вклад каждого из приложений в загруженность соты

В предыдущих разделах анализировались характеристики объёмов трафика без разбиения на приложения. Однако при прогнозировании загруженности соты полезно понимать, как распределяется нагрузка на соту по приложениям. На рисунках 3.11 – 3.13 представлены временные ряды количества уникальных пользователей, суммарного и среднего отправленного трафика. Для маркировки приложений используется прежняя цветовая легенда (см. рис. 2.6).

Ананлогичные графики можно построить как для полученного, так и для общего (сумма отправленного и полученного) трафика. Итого мы имеем девять 15-мерных (все типы приложений, кроме `Legacy Protocols`) временных рядов. На всех рядах отчетливо видны дневные паттерны, поэтому удачным решением для их прогнозирования будет применение моделей, описанных в разделе 2.4, в частности модели векторной авторегрессии с сезонными фиктивными переменными. Для каждого 15-мерного ряда обучим модель VAR(1) на нормированных данных до перерыва и вычислим метрику RMSE между прогнозом и нормированными данными после перерыва. Информация о точности прогнозирования характеристик каждого приложения по каждому ряду содержится в таблице 3.1.

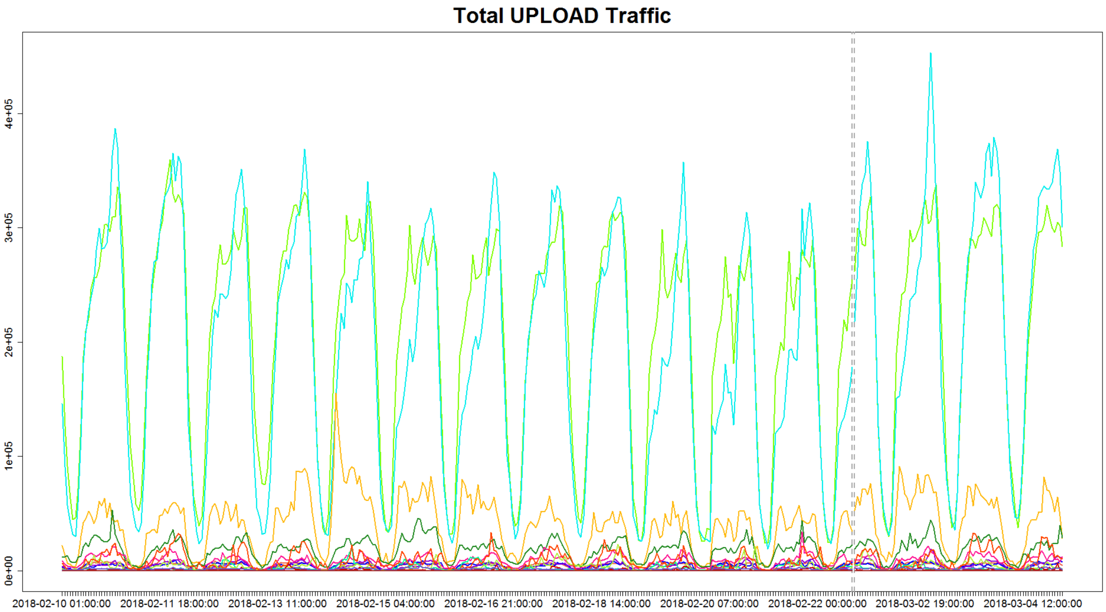



Рис. 3.11. Суммарный отправленный трафик по приложениям

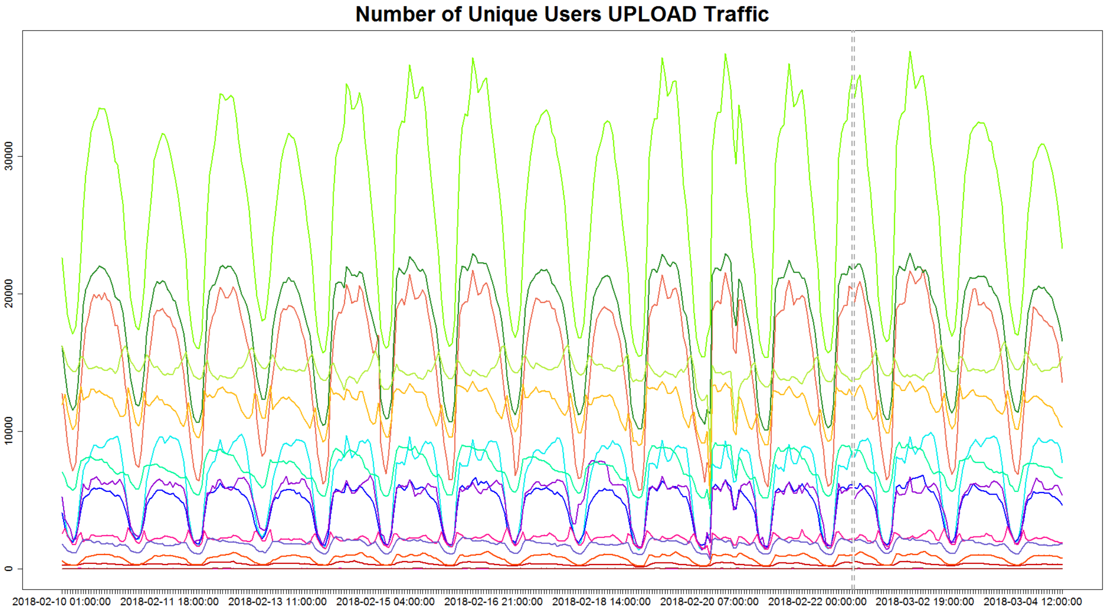



Рис. 3.12. Количество уникальных пользователей по отправленному трафику по приложениям

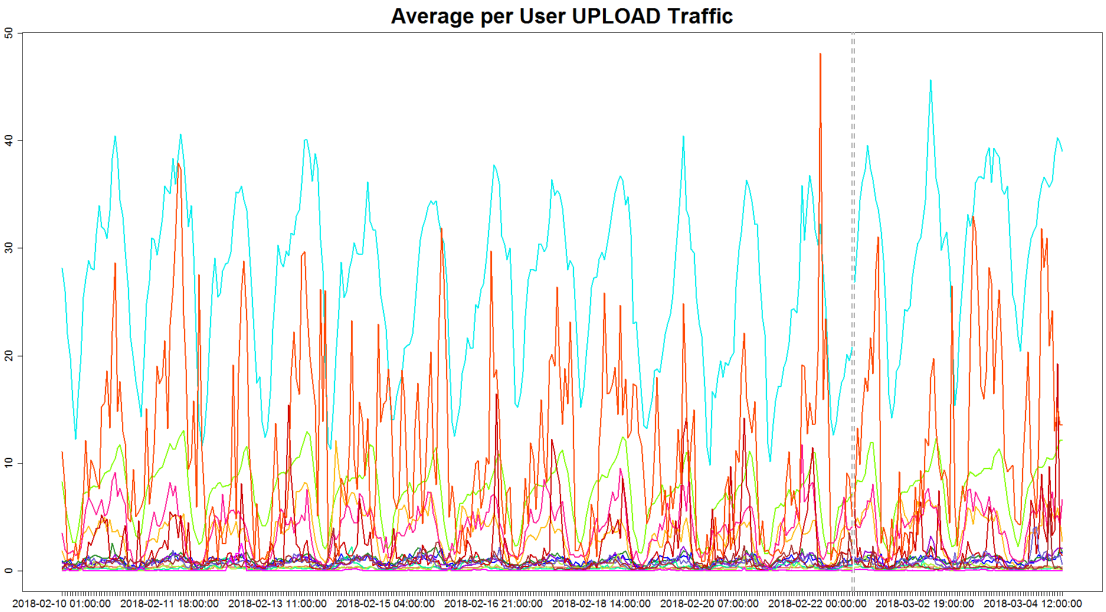



Рис. 3.13. Средний отправленный трафик по приложениям

| Applications | Upload | | | Download | | | Upload+Download | | |
|---|---|---|---|---|---|---|---|---|---|
| | Users | Total | Average | Users | Total | Average | Users | Total | Average |
| Web | 0.109 | 0.078 | 0.082 | 0.108 | 0.067 | 0.067 | 0.109 | 0.075 | 0.079 |
| Instant Messaging | 0.091 | 0.056 | 0.051 | 0.091 | 0.099 | 0.114 | 0.091 | 0.062 | 0.071 |
| Streaming | 0.082 | 0.137 | 0.168 | 0.081 | 0.154 | 0.179 | 0.082 | 0.137 | 0.168 |
| File Transfer | 0.072 | 0.094 | 0.091 | 0.072 | 0.119 | 0.11 | 0.072 | 0.094 | 0.091 |
| Security | 0.095 | 0.127 | 0.137 | 0.095 | 0.069 | 0.069 | 0.095 | 0.131 | 0.145 |
| Others | 0.1 | 0.086 | 0.096 | 0.1 | 0.126 | 0.143 | 0.1 | 0.082 | 0.09 |
| VoIP | 0.081 | 0.163 | 0.176 | 0.087 | 0.145 | 0.159 | 0.081 | 0.158 | 0.17 |
| Music Streaming | 0.102 | 0.086 | 0.103 | 0.102 | 0.087 | 0.095 | 0.102 | 0.086 | 0.103 |
| Mail | 0.103 | 0.088 | 0.087 | 0.103 | 0.077 | 0.078 | 0.103 | 0.096 | 0.094 |
| Games | 0.067 | 0.146 | 0.148 | 0.068 | 0.399 | 0.464 | 0.068 | 0.146 | 0.147 |
| Terminals | 0.12 | 0.16 | 0.175 | 0.119 | 0.216 | 0.194 | 0.119 | 0.17 | 0.181 |
| P2P | 0.106 | 0.211 | 0.174 | 0.108 | 0.162 | 0.144 | 0.106 | 0.188 | 0.17 |
| Network Operation | 0.087 | 0.244 | 0.289 | 0.088 | 0.093 | 0.09 | 0.089 | 0.262 | 0.303 |
| File Systems | 0.12 | 0.049 | 0.03 | 0.131 | 0.038 | 0.033 | 0.133 | 0.037 | 0.034 |
| DB Transactions | 0.204 | 0.193 | 0.208 | 0.184 | 0.068 | 0.052 | 0.205 | 0.114 | 0.101 |

Таблица 3.1. Ошибки прогнозирования характеристик каждого приложения по каждому ряду

# Заключение

Методы прогнозирования и анализа мобильного трафика, описанные в работе, позволяют тонко настроить объёмы выделяемых ресурсов при сегментации мобильной сети. Они применимы для самых различных агрегаций объёмов трафика: исследователь волен варьировать размер окна по времени (для представленных данных ширина окна обязана быть кратна одному часу, однако подходы могут успешно применяться и на других временных масштабах в случае доступности соответствующих наблюдений) и способ группировки приложений (рассматривать приложения отдельно, по кластерам или же анализировать полный трафик без разбиения на приложения). Аналогичные замечания справедливы и для статистической процедуры выявления аномальных наблюдений. Возможность столь гибкого анализа обусловлена выбором подходящего класса вероятностных распределений — обобщённого гамма-распределения, а также универсальных подходов к анализу временных рядов.

Дальнейшие исследования могут быть ориентированы на решение задач прогнозирования характеристик мобильной сети с использованием различных нейронных сетей [15, 16] в совокупности с моделями на основе обобщённых гамма-распределений для повышения точности прогнозов. Однако это более вычислительно сложные подходы по сравнению с рассмотренными в данной работе. Продолжением анализа аномальных наблюдений может служить составление профиля пользователя по количеству наблюдений, признанных аномальными, за определенный промежуток времени.

Несмотря на успешность использования обобщённого гамма-распределения для аппроксимации распределений объёмов трафика, стоит отметить, что у эмпирического распределения иногда возникает небольшой «горб» после пика у нуля, который обобщённое гамма-распределение не способно огибать. Особенно часто этот эффект наблюдается при анализе трафика без разбиения по приложениям (рисунки 2.1, 2.3). Естественным продолжением исследования в этом направлении является использование конечной смеси обобщённых гамма-распределений (на это также «намекает» кластеризация приложений — рисунок 2.13).

Результаты исследования были представлены на научной конференции «Ломоносовские чтения» и опубликованы в сборнике тезисов [17].

# Список литературы

[1] Симаков Д. В., Кучин А. А. Анализ статистических характеристик Интернет-трафика в магистральном канале // T-Comm: Телекоммуникации и транспорт. – 2015. – Том 9. – No 5. – С. 31–35.

[2] Исследование возможностей прогнозирования трафика сети мобильной связи / В. М. Безрук, И. В. Корсун, В. А. Тихонов, Н. В. Кудрявцева // Восточно-Европейский журнал передовых технологий. 2010. Vol. 4, Iss. 9 (46).

[3] Шелухин О. И., Сакалема Д. Ж., Филинова А. С. Обнаружение вторжений в компьютерные сети. Сетевые аномалии. – М.: Горячая линия – телеком, 2013. – 220 с.

[4] Терновой О. С., Шатохин А. С. Использование байесовского классификатора для получения обучающих выборок, позволяющих определять вредоносный трафик на коротких интервалах // Известия Алтайского государственного университета. 2013. Iss. 1-1 (77).

[5] Шелухин О. И., Филинова А. С. Обнаружение сетевых аномальных выбросов трафика методом разладки Бродского-Дарховского // T-Comm: Телекоммуникации и транспорт. 2013. Vol. 7, Iss. 10.

[6] Шелухин О. И., Судариков Р. А. Анализ информативных признаков в задачах обнаружения аномалий трафика статистиче-

скими методами // T-Comm: Телекоммуникации и транспорт. 2014. Vol. 8, Iss. 3.

[7] Stacy E. W. A Generalization of the Gamma Distribution // The Annals of Mathematical Statistics. 1962. Vol. 33, Iss. 3.

[8] Буткевич М. Н. Статистические характеристики и модели трафика мобильных приложений // Вестник ассоциации вузов туризма и сервиса. 2009. Iss. 1.

[9] Goode B. Voice over Internet protocol (VoIP) // Proceedings of the IEEE. 2002. Vol. 90, Iss. 9.

[10] Chernoff H., Lehmann E. L. The Use of Maximum Likelihood Estimates in $\chi^2$ Tests for Goodness of Fit // The Annals of Mathematical Statistics. 1954. Vol. 25, Iss. 3.

[11] Ward, Joe H. Hierarchical Grouping to Optimize an Objective Function // Journal of the American Statistical Association. 1963. Vol. 58, Iss. 301.

[12] Akaike H. A new look at the statistical model identification // IEEE Transactions on Automatic Control. 1974. Vol. 19, Iss. 6.

[13] Pfaff B., Stigler M. Package 'vars' [Электронный ресурс] // cran.r-project.org: The Comprehensive R Archive Network. URL: https://cran.r-project.org/web/packages/vars/vars.pdf (дата обращения: 10.05.2022).

[14] Korolev V. Yu., Gorshenin A. K. Probability models and statistical tests for extreme precipitation based on generalized negative binomial distributions // Mathematics. 2020. Vol. 8, Iss. 4. Art. No. 604.

[15] Mobile traffic forecasting for maximizing 5G network slicing resource utilization / V. Sciancalepore, K. Samdanis, X. Costa-Perez, D. Bega, M. Gramaglia, A. Banchs // IEEE INFOCOM. 2017. P. 1–9.

[16] Probabilistic Forecasting of Sensory Data With Generative Adversarial Networks - ForGAN / A. Koochali, P. Schichtel, A. Dengel, S. Ahmed // IEEE Access. 2019. Vol. 7. P. 63868–63880.

[17] Горбунов С. А., Горшенин А. К. Об обобщённом гамма-распределении в задачах анализа мобильного трафика // Научная конференция «Ломоносовские чтения»: тезисы докладов. 14–22 апреля 2022 года. Секция Вычислительной математики и кибернетики. 2022. С. 166–167.