УДК 004.75

Цыберный Н.Н.,
студент магистратуры
факультет «Информатика и системы управления»
ФГБОУ ВО Московский государственный технический
университет имени Н.Э. Баумана
Россия, г. Москва

ОБНАРУЖЕНИЕ АНОМАЛИЙ СЕТЕВОГО ТРАФИКА В ПРОЦЕССЕ МОНИТОРИНГА ИНФРАСТРУКТУРЫ КОРПОРАТИВНОЙ СЕТИ

Аннотация: *Рассмотрена возможность мониторинга сетевой инфраструктуры путём обнаружения аномальных значений загрузки канала передачи данных. Предложен метод построения профиля нормального функционирования канала с использованием статистических моделей ARMA и ARIMA. В заключительной части работы выполнена имитация процесса мониторинга и продемонстрирована его применимость.*

Ключевые слова: *система мониторинга, аномалии, нестационарные процессы, модель ARIMA.*

Abstract: *The possibility of monitoring the network infrastructure by detecting abnormal values of the data link load has been considered. A method is proposed for constructing a channel normal functioning profile using statistical models ARMA and ARIMA. In the final part of the work, an imitation of the monitoring process was performed and its applicability was demonstrated.*

Key words: *monitoring system, anomalies, non-stationary processes, ARIMA model.*

С ростом автоматизации бизнес-процессов количество проходящего через корпоративную сеть трафика увеличивается, а его структура усложняется.

Помимо сообщений почтовых клиентов и веб-страниц генерируется большое количество критически важных для бизнеса данных - начиная от персональных и заканчивая данными платёжных систем. Их потеря может привести к существенным финансовым проблема или испортить репутацию компании среди партнёров и клиентов. В связи с этим возникает острая необходимость обеспечения высокой доступности, а также отказоустойчивости ИС и сети передачи, участвующей в процессе обмена информацией.

Решить данную задачу позволяет система мониторинга, своевременно оповещающая о сбоях, позволяющая проводить комплексный анализ работы сети, дающая подробную картину функционирования и производительности её элементов – объектов мониторинга [1, с.31].

Система мониторинга осуществляет сбор и обработку данных о контролируемом объекте по протоколу SNMP, что позволяет выдвинуть в большинстве случаев два суждения о его состоянии – доступен или недоступен. При данном методе контроля отследить деградацию или перегрузку сервиса, вызванную сбоями в работе программного обеспечения, сетевого оборудования или сегментов сети, находящихся в чужой зоне ответственности, становится затруднительно. В таком случае целесообразно применять методы мониторинга, выполняющие анализ косвенных признаков состояния объекта.

В качестве таких признаков в данной работе рассматривается объём данных переданных через канал связи. Обнаружение аномальных значений данного параметра позволит своевременно выявить генерирующий их источник, оценить его состояние, степень влияния на инфраструктуру и устранить причину сбоя. Под аномальными понимаются ранее не наблюдаемые значения или их резкие изменения за определённый временной интервал. Техника обнаружения аномальных значений построена на сопоставлении текущего состояния сетевой инфраструктуры с некими определенными заранее признаками, характеризующими штатное функционирование сетевой инфраструктуры [4, с. 81]. Основным этапом работы является формирование на основе собранной

информации о загрузке канала его профиля или модели, отображающей нормальный режим функционирования.

Предлагаемый метод мониторинга включает прогнозирующую и обнаруживающую фазы. Предполагается, что собранные данные отражают нормальный режим функционирования корпоративной сети и по ним выполняется прогноз значении контролируемого параметра на несколько шагов вперёд. При поступлении новых данных, значительно отличающихся от спрогнозированных, выдвигается гипотеза об их аномальности. Ключевым этапом в данном методе является построение прогнозирующей модели.

Измеренные значения скорости передачи данных x_i в канале представим в виде временного ряда $Y = y(1), y(2), y(3), \dots, y(n)$. Наиболее распространённой и эффективной для анализа временных рядов является статистическая модель авторегрессии-скользящего среднего (Autoregressive Moving Average) и интегрированная модель авторегрессии-скользящего среднего (Autoregressive Integrated Moving Average).

В общем виде модель авторегрессии-скользящего среднего обозначается как ARMA(p,q), где p−порядок авторегрессии, q−порядок скользящего среднего. Её можно представить, как:

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q},$$ где

y_t – значение временного ряда при определённом времени t;

$y_{t-1}, y_{t-2}, \dots, y_{t-p}$ – значения ряда в моменты времени $t-1, t-2, \dots, t-p$;

$\phi_0, \phi_1, \phi_2, \dots, \phi_p$; $\theta_0, \theta_1, \theta_2, \dots, \theta_q$ – подлежащие оценки параметры модели;

$\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_{t-q}$ – значения шума в q периоде времени назад;

Данная модель учитывает зависимость текущих значений ряда от самих себя в прошлом, а также зависимости настоящего и прошлых значений шума. Модель имеет одно существенное ограничение, заключающееся в том, что она описывает только стационарные процессы. В широком смысле под

стационарностью ряда можно понимать отсутствие в нём тренда, сезонности и циклов.

Большинство реальных физических процессов не являются стационарными и имеют изменяющиеся во времени характеристики. Для приведения их к стационарности существуют различные методики, одной из которых является взятие конечное число раз разностей членов ряда и формирования нового ряда $\Delta Yt = Yi(t+1) - Y(i)t$.

В итоге, если при взятии разности последовательно d раз ряд приводится к стационарному, то для его описания можно использовать интегрированную модель авторегрессии-скользящего среднего, обозначающуюся как ARIMA(p,d,q). В данной модели параметр d соответствует количеству взятых разностей и показывает порядок интегрирования I(d) ряда, в остальном модель схожа с ARMA(p,q).

Прежде чем сделать окончательный выбор в пользу одной из моделей необходимо дать оценку стационарности исследуемого процесса. Помимо визуального исследования ряда на наличие тренда или периодических колебаний, может быть применён формальный тест Дики-Фуллера (Dickey-Fuller unit-root test) [6, с. 427].

В данной работе используются данные, собранные с внешнего канала связи реальной сетевой инфраструктуры компании. Они представляют из себя полученные через каждые 15 секунд значения SNMP – счётчика порта, содержащего количество бит переданной информации V через канал связи. Временная агрегация Δt составила 15, 60 и 300 сек. Таким образом для трёх различных величин Δt были сформированы агрегированные ряды $Y(i) = \nu(t_1), \nu(t_2), .., \nu(t_n)$ на временном интервале T (28 дней), где:

$$\nu(t_n) = \frac{V_{t_n} - V_{t_n-1}}{\Delta t}$$

Их длинна составила 161280, 40320 и 8064 значений соответственно. Фрагменты рядов показаны на рисунке 1.

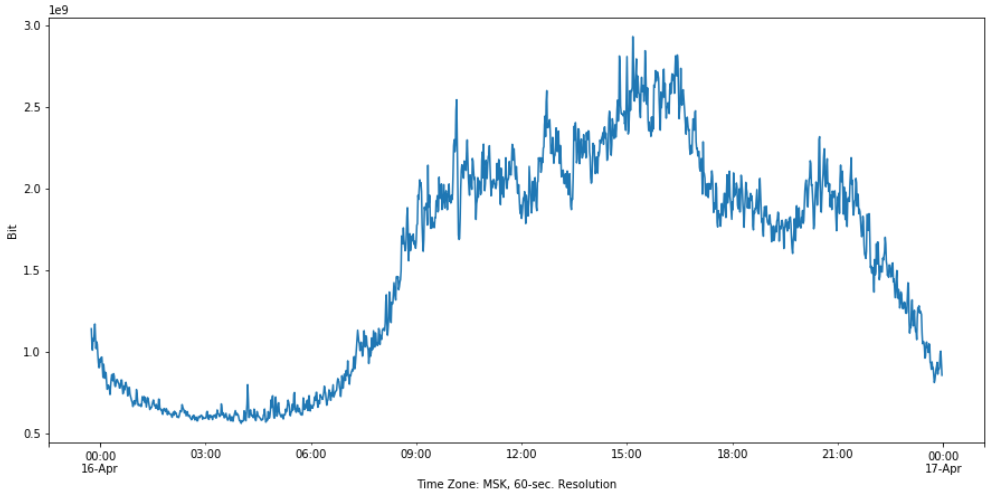

Рисунок 1. Фрагмент ряд длиной в 24 часа при $\Delta t = 60$ секунд

Перед построением модели выполняется процесс предобработки, целью которого является восстановление целостности и сглаживание всплесков в данных. Это необходимо для того, чтобы устранить пропущенные значения и снизить влияние выбросов на процесс моделирования, при этом пустые поля заполняются методом интерполяции, а затем применяется сглаживание скользящей средней с окном в 5 значений.

Процесс оценки стационарности состоит из следующих шагов:

1. Устанавливаются начальные значения ширины окна w и шага n
2. Из исходного временного ряда $Y(i)$ длинной l формируется тестовый временной ряд $Y(ij)$, включающий точки от t_0 до t_w
3. Для ряда $Y(ij)$ вычисляется значение ADF - теста. При уровне значимости теста p-value > 0,05 ряд признаётся нестационарным, в обратном случае – стационарным.
4. Окно сдвигается на n точек вперёд, формируется следующий тестовый временной ряд $Y(i(j+1))$, включающий точки от $t_{(j\cdot 1)\cdot n}$ до $t_{w+(j\cdot n)}$ и выполняется пункт 3. Алгоритм останавливается при $t_{w+(j\cdot n)} = t_l$ или $t_{w+(j\cdot n)} > t_l$, тогда ряд $Y(ij)$ не учитывается.

Зависимость количества нестационарных рядов от w и n , выраженная в % от общей суммы тестовых временных рядов Y(ij) показана на рисунках 2 -3.

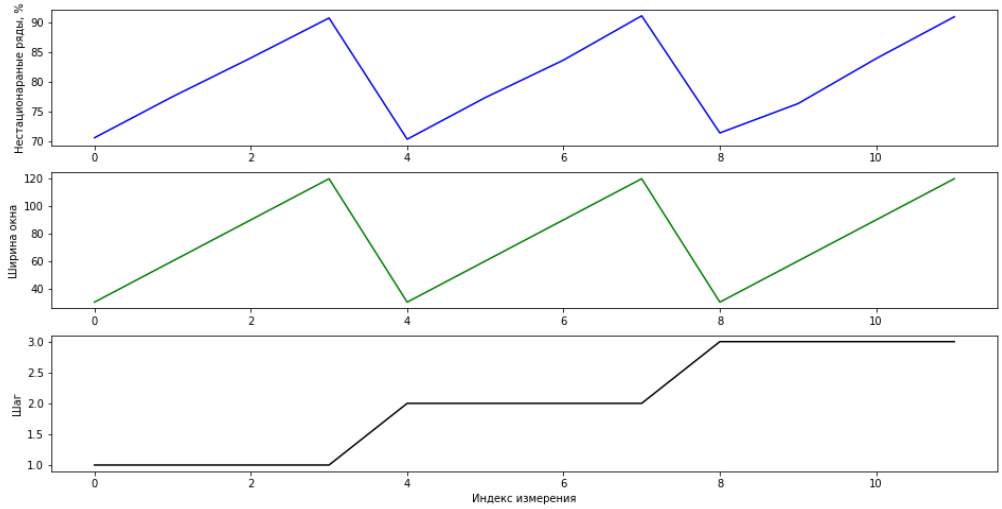

Рисунок 2. Зависимость количества нестационарных рядов от общей суммы тестовых временных рядов $Y(ij)$ при $\Delta t = 300$ секунд

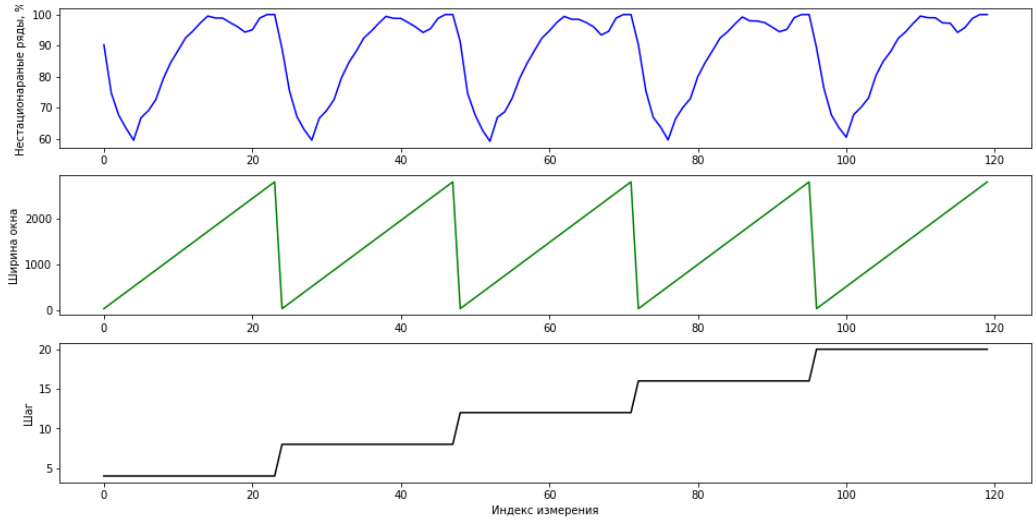

Рисунок 3. Зависимость количества нестационарных рядов от общей суммы тестовых временных рядов $Y(ij)$ при $\Delta t = 15$ секунд

Из приведённых графиков видно, что стационарность выборки не зависит от шага окна n и при разной временной агрегации Δt имеет одинаковую форму зависимости от длины исследуемого ряда. Это свойство объясняется фрактальностью, т.е. самоподобием сетевого трафика. Наиболее общим определением фрактального процесса является его неформальное определение как случайного процесса, статистические характеристики которого проявляют

свойства масштабирования. Самоподобный процесс существенно не меняет вида при рассмотрении в различных масштабах по шкале времени. [3, с. 97]. С ростом длины выборки возрастает количество нестационарных рядов и при её значении в 600 точек или 10 часов при окне агрегирования в 60 секунд их доля достигает 91 %. При визуальном исследовании ряда данной длины видно (рисунок 4), что нестационарность процесса обусловлена наличием тренда, имеющего ярко выраженный характер.

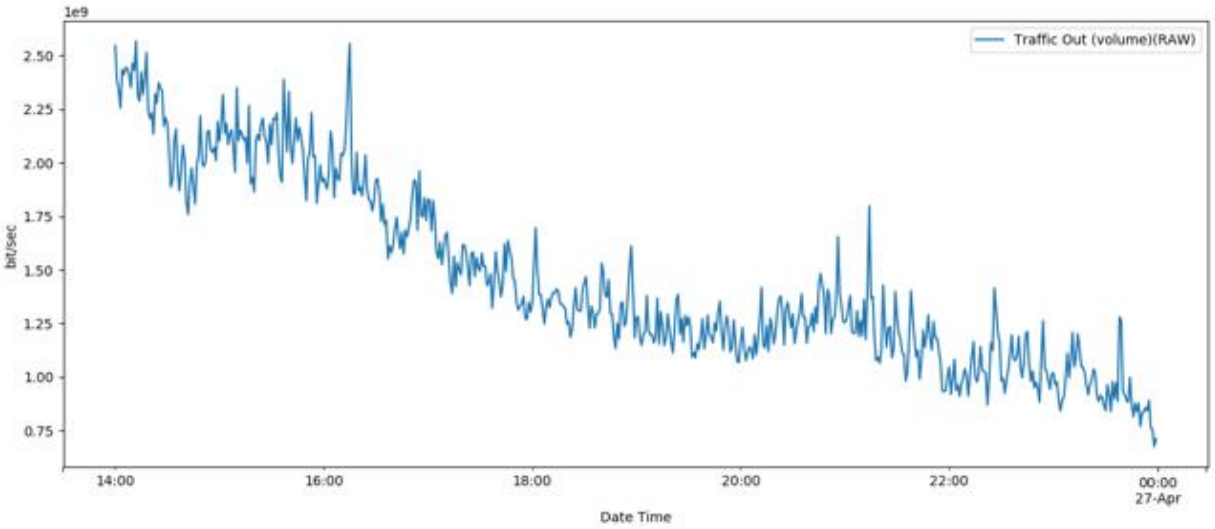

Рисунок 4. Нестационарный ряд из 600 точек (10 часов) при окне агрегирования в 60 секунд

Так как предполагается выполнять оперативный мониторинг в человеческих масштабах восприятия времени и принятых в компании временных нормативах об оповещении, горизонт прогнозирования (шаг окна) будет ограничен 5 минутами, что для имеющихся рядов с дискретизацией 15, 60 и 300 секунд составит 20, 5 и 1 точку соответственно при этом прогноз будет являться краткосрочным. Порядок интегрирования I(n) для каждого окна будет вычисляться на основе ADF-теста и перехода к интегрированному ряду $\Delta Y(i)t = Yi(t+1) - Y(i)t$ до тех пор, пока ряд не будет приведён к стационарному, на практике редко превышает I(2). При увеличении порядка интегрирования растёт дисперсия прогноза, поэтому глубина истории должна обеспечивать

наименьший процент нестационарных выборок на исходном временном ряде с одной стороны и требуемую точность прогноза, с другой стороны.

Выявление компонент (тренда, сезонности, цикличности, случайной компоненты и остатка) всей имеющейся исследуемой выборки проводилось визуальным анализом STL (Seasonal and Trend decomposition using Loess) декомпозиции. STL, в отличие от SEATS и X11 методов, позволяет выявить любой тип сезонности, не только ежемесячные и квартальные. Так же данный метод является устойчивым к выбросам и минимизирует их влияние на оценку компонентов [8].

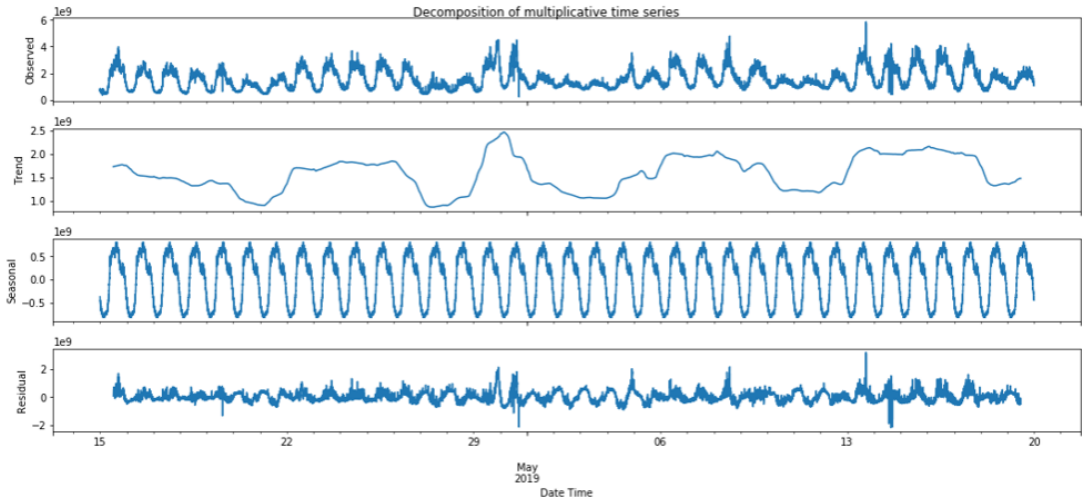

Рисунок 5. STL декомпозиция временного ряда значений скорости передачи данных

По графикам временного ряда на рисунке 5, полученным после декомпозиции, можно сделать вывод, что имеются ярко выраженные тренды в утренние и вечерние часы, а также суточная сезонность. Помимо неё прослеживаются и недельные изменения, при которых значения ряда снижаются к вечеру пятницы и остаются относительно низкими вплоть до утра понедельника. Такая динамика обусловлена сокращением пользовательской активности в ночные и нерабочие часы по будням и выходные дни.

Поскольку при визуальном анализе всей имеющейся выборки было замечено, что профиль каждого дня повторяется от недели к неделе был выбран подход ежедневного профилирования трафика. Для учёта незначительных

отклонений в одном и том же дне разных недель тренировочный участок был сформирован из средних значений двух следующих подряд понедельников и вторников (рисунок 6).

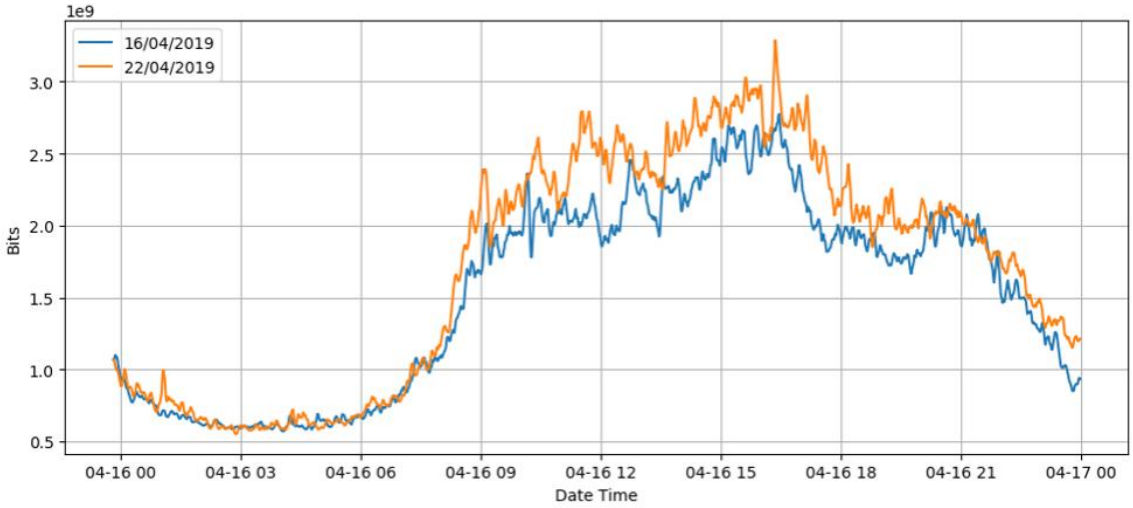

Рисунок 6. Тренировочный участок

Третья и четвёртая недели были исключены из-за аномального профиля в следствии аварийных событий в этот день. Значения этих дней были использованы для формирования тестовых участков.

Модель строится на основе методологии Бокса-Дженкинса, процесс описан в [7, с. 78]. Имитацию процесса мониторинга предлагается выполнить по следующим этапам:

1. Из всего диапазона значений ряда $Y(i)$ выбирается тренировочный участок. На нём из окна длинной l значений формируется новый ряд $Y(ij)$.

2. Ряд $Y(ij)$ подаётся на вход функции оценки параметров. Параметры p, d, q методом перебора подставляются в модель ARIMA (p, d, q) и вычисляется информационный критерий AIC. Функция возвращает модель с параметрами, обеспечивающими минимальный AIC.

3. По полученной модели строится прогноз значений на n точек вперёд

4. Для интервала от $(j+1)\cdot l$ до $(j+1)\cdot l+n$ точек исходного временного ряда и прогнозных n точек вычисляется абсолютный средний процент отклонения (MAPE).

5. Окно сдвигается на n точек вперёд, формируется следующий тестовый временной ряд $Y(i(j+1))$ и выполняются пункты 2,3,4,5. Алгоритм останавливается при достижении окном конца исходного ряда.

В качестве параметра для оценки расхождения наблюдаемых значений от профиля воспользуемся средней абсолютной ошибкой (mean percentage absolute error, MAPE). Она показывает насколько большими являются ошибки прогнозирования в сравнении с действительными значениями ряда и рассчитываемая как:

$$\frac{1}{n}\sum_{t=1}^{n}\frac{|Y_t - \hat{Y_t}|}{Y_t}$$

Где Y_t – фактические значения, $\hat{Y_t}$ - прогнозные значения для n точек ряда.

Используемый формальный критерий MAPE, на основе которого регистрируются изменения, является варьируемым параметром и в каждом случае должен подбираться индивидуально.

Так как мы тестируем последовательно участки из 5 точек примем, что:

- при увеличении MAPE на 5% для двух последовательных участков фиксируется плавный рост трафика. Аномалией не считается.
- при увеличении MAPE на 10% для двух последовательных участков фиксируется выброс в значениях.
- при снижении MAPE на 10% и более для двух последовательных участков фиксируется провал в значениях.
- если для двух участков после выброса или провала не наблюдается снижение оценочного параметра более чем на 10%, то имеет место сдвиг в значениях.

Для уменьшения ложных срабатываний, в период с 8:00 до 20:00, точность может быть снижена, т.к. эти часы являются рабочими и в них наблюдаются наибольшие колебания легитимного трафика, вызванные активностью пользователей.

На рисунке 7 показаны результаты эксперимента. Красными точками выделены значения, которые были признаны аномальными в сравнении с профилем, показанным зелёным цветом.

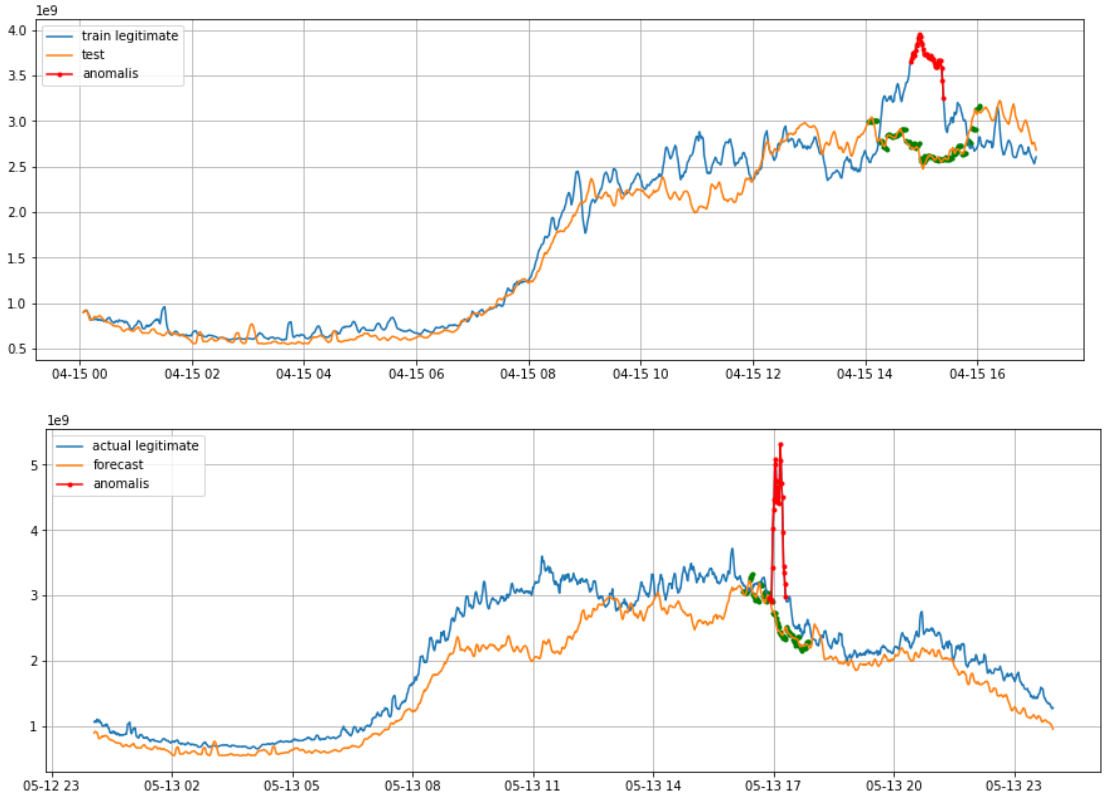

Рисунок 7. Выявленные аномалии в тестовых рядах

По результатам эксперимента сделан вывод, что предложенный метод может быть применён в процессе мониторинга корпоративной сетевой инфраструктуры. Наиболее продолжительным этапом в его реализации является процесс сбора данных для профилирования нормального режима функционирования канала связи. Но в сравнении с использованием нейронный сетей, требующих длительного обучения на тех же данных, построение статистических моделей занимает меньше времени. эффективность напрямую зависит от качаства данных, используемых для построения профиля, а также от установки оценочного параметра.

ИСПОЛЬЗОВАННЫЕ ИСТОЧНИКИ

1. В.В. Кореньков, В.В. Мицын, П.В. Дмитриенко. Архитектура системы мониторинга центрального информационно-вычислительного комплекса ОИЯИ// Информационные технологии и вычислительные системы. - 2012. - № 3. - С. 31-42.

2. С.А. Коноваленко, И.Д. Королев, Д.А. Новоселов. Базовые функциональные возможности существующих систем мониторинга вычислительных сетей// Приволжский научный вестник. - 2016. - № 12-1. - С. 65-70.

3. Н.Г. Треногин, Д.Е. Соколов. Фрактальные свойства потоков событий прикладного уровня в информационных системах // Вестник СибГАУ. 2017. Т. 4, № 1. С. 97–103.

4. Шелухин О.И., Сакалема Д.Ж., Филинова А.С.. Обнаружение вторжений и компьютерные сети., / О.И. Шелухин — М.: Горячая линия-Телеком, 2013. — 220с.

5. Cleveland, R.B., Cleveland, W.S., McRae, J.E., & Terpenning, I. J. (1990). STL: A seasonal-trend decomposition procedure based on loess. Journal of Official Statistics, *6*(1), 3–73.

6. Dickey, D.A., Fuller, W.A. (1979); Distribution of the estimators for autoregressive time series with a unit root, Journal of the American Statistical Association 74, 427--431.

7. Hyndman, Robin John; Athanasopoulos, George. / Forecasting: Principles and Practice**.** 2nd ed. OTexts, 2018. 384 p.

8. Электронный курс по анализу временных рядов: личная страница Brant Deppa, Ph.D. [Электронный ресурс]. URL: http://course1.winona.edu/bdeppa/FIN%20335/Handouts/Time_Series_Decomposition.html#stldecomposition (дата обращения 07.05.2019).