УДК 656.13.08

М.Г. Черевастов[1], Ю.И. Молев[2]

ПОЛУЧЕНИЕ АДДИТИВНОЙ МОДЕЛИ ВРЕМЕННОГО РЯДА, ХАРАКТЕРИЗУЮЩЕГО ДОРОЖНО-ТРАНСПОРТНУЮ АВАРИЙНОСТЬ НА ПРИМЕРЕ НИЖЕГОРОДСКОЙ ОБЛАСТИ

[1]Нижегородский государственный педагогический университет им. К. Минина
[2]Нижегородский государственный технический университет им. Р.Е. Алексеева

Дорожно-транспортная аварийность рассмотрена как упорядоченный по времени ряд (временной ряд). Для описания временного ряда выбрана аддитивная модель. В результате анализа ряда и аналитического выравнивания данных выделена общая динамика (тренд) его изменения, уравнение которой представляет собой полином второй степени (парабола). Проведено численное сглаживание ряда. На основе автокорреляционного анализа временного ряда установлена нелинейность тренда и наличие периодической компоненты с циклом в семь дней. Определено практическое отсутствие связи между уровнями ряда и моментом времени (сутки). Вычислены значения сезонной компоненты и построена модель временного ряда.

Ключевые слова: безопасность дорожного движения, дорожно-транспортная аварийность, временной ряд.

На основе методов математической статистики авторами ранее были исследованы официальные данные о количестве произошедших дорожно-транспортных происшествий, в которых погибли или получили ранения различные участники дорожного движения, а также был причинен материальный ущерб [1]. Были получены оценки числовых характеристик и параметров распределения случайных величин, а также установлены законы их распределения. С другой стороны, ежедневные статистические данные об общем количестве дорожно-транспортных происшествий [2] на территории Нижегородской области в 2016 г. можно представить как упорядоченную по времени информацию для формирования временного ряда $\{y_t\}$, общий анализ которого позволяет определить его природу и спрогнозировать его будущие значения [3].

Для настоящего исследования сформируем временной ряд и представим его графическим способом. Для данного случая длина ряда составляет 366 элементов – по числу дней в 2016 году. Каждый элемент ряда содержит в себе значение показателя уровня ряда – количество дорожно-транспортных происшествий, и период времени – сутки. Условимся, что уровни временного ряда мы будем обозначать y_t, а соответствующие им моменты времени через t. Таким образом, учитывая длину ряда, можно записать $t = 1,2,3,...,n$, а $n = 366$, т.е., значение показателя уровня ряда y_1 соответствует количеству происшествий 1 января 2016 г., а y_{32} соответствует происшествиям, зарегистрированным за сутки 1 февраля того же года. Для проведения анализа временного ряда нами использована аддитивная модель, широко применяемая для описания различных технических и экономических процессов. Согласно этой модели, значения уровней ряда определяются следующим выражением:

$$y_t = \widehat{y}_t + s_t + c_t + e_t \qquad (1)$$

где $\widehat{y}_t$ – тенденция (тренд), систематическая устойчивая долговременная динамика;

s_t – сезонная систематическая (периодическая) компонента, проявляющаяся на протяжении года;

c_t – циклическая систематическая (периодическая) компонента, описывающая длительные периоды времени (более одного года);

e_t – случайная нерегулярная компонента.

На рис. 1 изображен график исследуемого временного ряда. По оси абсцисс отложены моменты времени, а по оси ординат − число дорожно-транспортных происшествий, соответствующее этим моментам. На графике отчетливо видны колебания, повторяющиеся во времени, при этом большие значения уровни ряда принимают в начальные и конечные моменты времени.

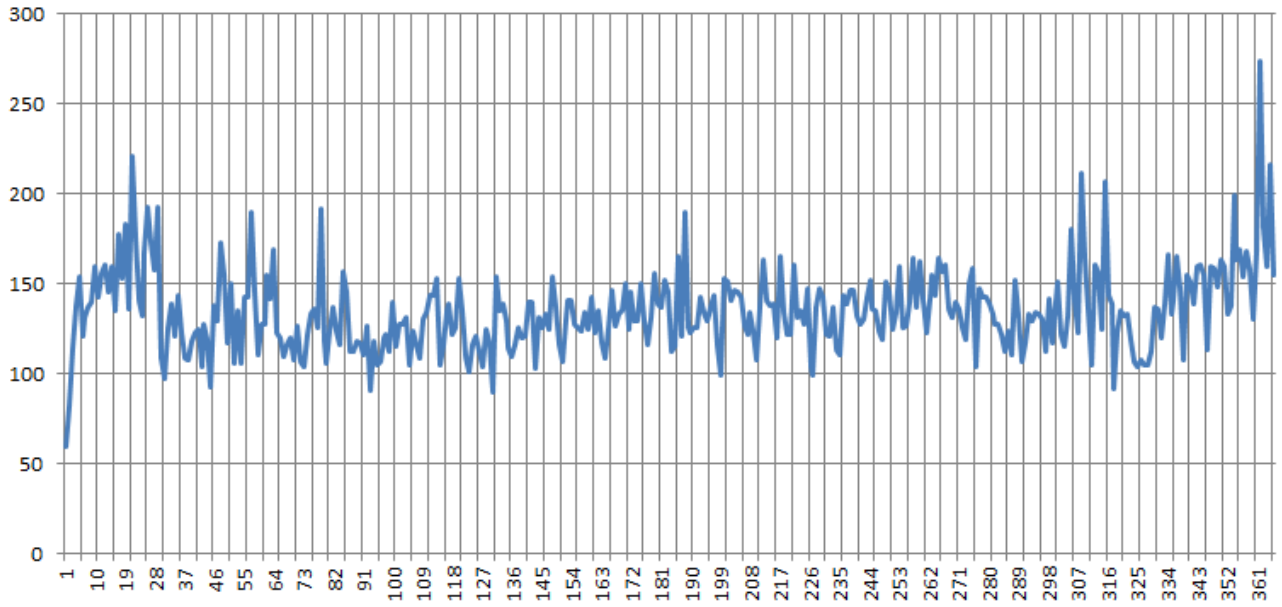

Рис. 1. График временного ряда, отражающий состояние дорожно-транспортной аварийности на территории Нижегородской области

Представив исследуемый нами временной ряд графически, проверим его на стационарность. Так мы сможем определить наличие или отсутствие тенденции (тренда) в изучаемых данных. Проверим гипотезу H_0 о случайности тенденции (тренда) ряда при альтернативной гипотезе H_1 о не случайности тренда. Для этого разобьем временной ряд на две равные половины и сравним их средние значения, рассчитав по выражению (2) статистику критерия Стьюдента:

$$t_{расч} = \frac{\overline{y_I} - \overline{y_{II}}}{\sqrt{(n_I - 1)d_I + (n_{II} - 1)d_{II}}} \cdot \sqrt{\frac{n_I n_{II}(n_I + n_{II} - 2)}{n_I + n_{II}}} \qquad (2)$$

где n_I и n_{II} – длины первой и второй частей ряда ($n_I = n_{II} = 183$);

$\overline{y}_I$ и $\overline{y}_{II}$ – средние значения первой и второй половин ряда;

d_I и d_{II} – оценки дисперсий первой и второй половин ряда.

Средние значения и оценки дисперсий определим по следующим выражениям:

$$\bar{y}_I = \frac{1}{n_I}\sum_{t=1}^{183} y_t \tag{3}$$

$$\bar{y}_{II} = \frac{1}{n_{II}}\sum_{t=184}^{366} y_t \tag{4}$$

$$d_I = \frac{1}{n_I - 1}\sum_{t=1}^{183}\left(y_t - \bar{y}_I\right)^2 \tag{5}$$

$$d_{II} = \frac{1}{n_{II} - 1}\sum_{t=184}^{366}\left(y_t - \bar{y}_{II}\right)^2 \tag{6}$$

Результаты расчета числовых характеристик, вычисленных по формулам (3)–(6), сведены в табл. 1.

Таблица 1.
Значения числовых характеристик

Часть ряда	Оценка математического ожидания	Оценка дисперсии
I	$\bar{y}_I = 130{,}67$	$d_I = 475{,}46$
II	$\bar{y}_{II} = 138{,}89$	$d_{II} = 537{,}08$

Подставляя данные из табл. 1 в (2), получаем расчетную величину $t_{рас}$:

$$t_{расч} = -3{,}49$$

Сравнивая полученное значение критерия по абсолютной величине с критическим значением $t_{кр}$ распределения Стьюдента [4], определенным для уровня значимости $\alpha = 0{,}05$ и числа степеней свободы $df = n_I + n_{II} - 2 = 364$, сделаем вывод о принятии или отвержении гипотезы о случайности тренда нашего временного ряда. Имеем:

$$t_{кр} = 1{,}97$$

$$\left|t_{расч}\right| = 3{,}49 > 1{,}97$$

В данном случае, учитывая неравенство, нами отвергается гипотеза H_0 о случайности тренда временного ряда и принимается альтернативная гипотеза H_1 о не случайности тенденции временного ряда, т.е. тренд присутствует, а сам изучаемый временной ряд не стационарен.

Далее, для устранения случайных колебаний уровней временного ряда и выделения его тенденции проведем численное сглаживание ряда методом простого скользящего среднего [3]. Сглаженные значения уровней по пяти элементам ряда определяются следующим образом:

$$\tilde{y}_t = \frac{1}{5}\left(y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2}\right) \tag{7}$$

где $\tilde{y}_t$ – сглаженное значение соответствующего уровня ряда.

Вместе с тем, необходимо было определить сглаженные значения двух первых и двух последних элементов временного ряда, для чего применены следующие выражения [3]:

$$\tilde{y}_1 = \frac{1}{5}\left(3y_1 + 2y_2 + y_3 - y_5\right) \tag{8}$$

$$\tilde{y}_2 = \frac{1}{10}\left(4y_1 + 3y_2 + 2y_3 + y_5\right) \quad (9)$$

$$\tilde{y}_{365} = \frac{1}{10}\left(y_{366} + 3y_{365} + 2y_{364} + y_{362}\right) \quad (10)$$

$$\tilde{y}_{366} = \frac{1}{5}\left(3y_{366} + 2y_{365} + y_{364} - y_{362}\right) \quad (11)$$

Сглаженный временной ряд изображен на рис. 2 в виде кривой, обозначенной красным цветом. Как видно из графика, тенденция изменения уровней ряда носит явно нелинейный характер. В дальнейшем сглаженный ряд нам понадобится для выделения сезонной компоненты – s_t.

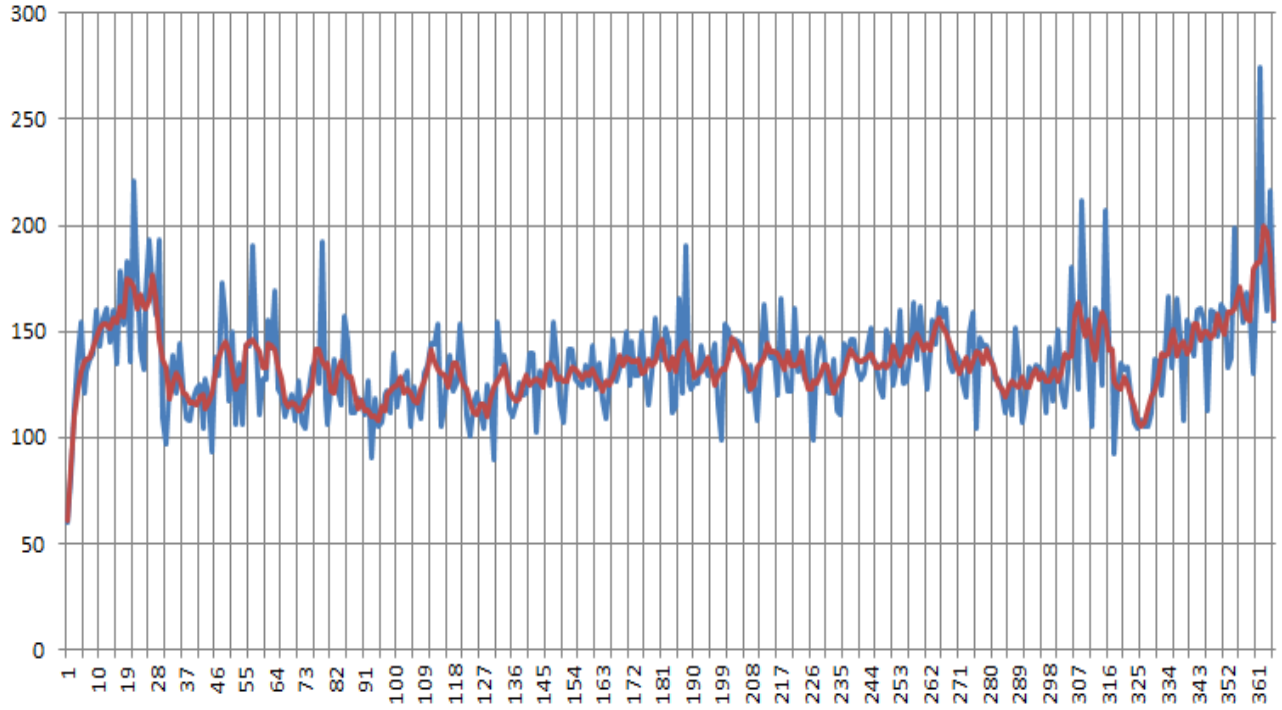

Рис. 2. Сглаживание временного ряда методом простого скользящего среднего

На следующем этапе исследования проведем автокорреляционный анализ временного ряда для установления его структуры (природы), подтверждения линейности (нелинейности) тренда, выявления периодических компонент. Для этого вычислим по формуле (12) коэффициенты автокорреляции различных порядков от первого до четырнадцатого.

$$r_l = \frac{\sum_{t=l+1}^{n}\left(y_t - \bar{y}_t\right)\left(y_{t-l} - \bar{y}_{t-l}\right)}{\sqrt{\sum_{t=l+1}^{n}\left(y_t - \bar{y}_t\right)^2 \cdot \sum_{t=l+1}^{n}\left(y_{t-l} - \bar{y}_{t-l}\right)^2}} \quad (12)$$

где $\bar{y}_t = \frac{1}{n-l}\sum_{t=l+1}^{n} y_t$ и $\bar{y}_{t-l} = \frac{1}{n-l}\sum_{t=l+1}^{n} y_{t-l}$.

Результаты расчетов коэффициентов автокорреляции приведены в табл. 2.

Таблица 2.
Значения коэффициентов автокорреляции

l	r_l	l	r_l
1	0,353361777	8	0,257686836
2	0,206623135	9	0,099934806
3	0,26141857	10	0,04788744
4	0,270496658	11	0,065501051
5	0,142518731	12	0,030629521
6	0,190029689	13	0,040592433
7	0,384957854	14	0,161552431

Теперь по имеющимся данным (табл. 2) построим коррелограмму:

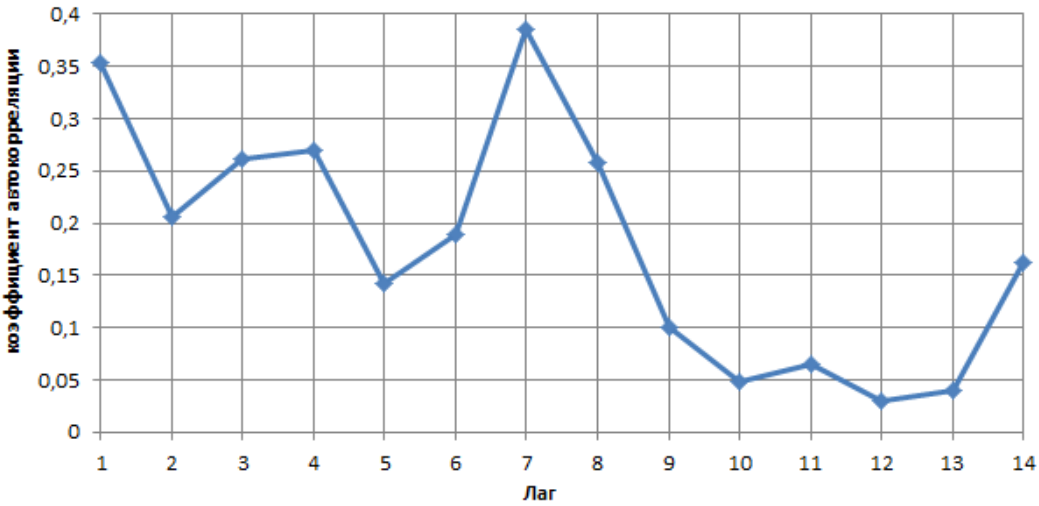

Рис. 3. Коррелограмма

Проверка полученных коэффициентов автокорреляции на статистическую значимость [3] показала, что, начиная с первого порядка по девятый, а также четырнадцатый, коэффициенты в целом статистически значимы и превышают критические значения коэффициентов корреляции Пирсона, рассчитанных для уровня значимости $\alpha = 0{,}05$ и соответствующего числа степеней свободы. Таким образом, по величинам $r_1 = 0{,}35$ и $r_7 = 0{,}38$ можно сделать вывод о том, что временной ряд имеет в своем составе как трендовую составляющую, так и периодическую сезонную, причем тренд обладает нелинейной тенденцией $(r_1 < 0{,}7)$. Ряду свойственна цикличность с периодом в семь моментов времени (коэффициент автокорреляции с лагом 7 имеет наибольшее значение), т.е., недельный цикл на ежедневных данных.

Теперь, проведя автокорреляционный анализ ряда, перейдем к составлению уравнения общей динамики (тренда) изучаемого временного ряда. Для этого произведем аналитическое выравнивание статистических данных, применяя метод наименьших квадратов. Как было отмечено нами ранее, тренд носит нелинейную тенденцию. Следовательно, для его моделирования будет применена полиномиальная (квадратичная) функция, точность использования которой будет оценена ниже.

Таким образом, уравнение тренда имеет следующий вид:

$$\hat{y}_t = at^2 + bt + c \tag{13}$$

где a, b, c – постоянные коэффициенты.

Определить постоянные коэффициенты, можно используя метод наименьших квадратов [5], решив систему уравнений (14):

$$\begin{cases} a\sum\limits_{t=1}^{366} t^4 + b\sum\limits_{t=2}^{366} t^3 + c\sum\limits_{t=1}^{366} t^2 = \sum\limits_{t=1}^{366} t^2 y_t \\ a\sum\limits_{t=1}^{366} t^3 + b\sum\limits_{t=1}^{366} t^2 + c\sum\limits_{t=1}^{366} t = \sum\limits_{t=1}^{366} t y_t \\ a\sum\limits_{t=1}^{366} t^2 + b\sum\limits_{t=1}^{366} t + 366c = \sum\limits_{t=1}^{366} y_t \end{cases} \tag{14}$$

В результате получаем: $a = 0{,}0006$, $b = -0{,}1751$ и $c = 140{,}63$.

Соответственно уравнение тренда имеет вид:

$$\hat{y}_t = 0{,}0006t^2 - 0{,}1751t + 140{,}63 \tag{15}$$

Представим полученное уравнение тренда графическим способом. На рис. 4 красным цветом изображена кривая, являющаяся общей тенденцией временного ряда.

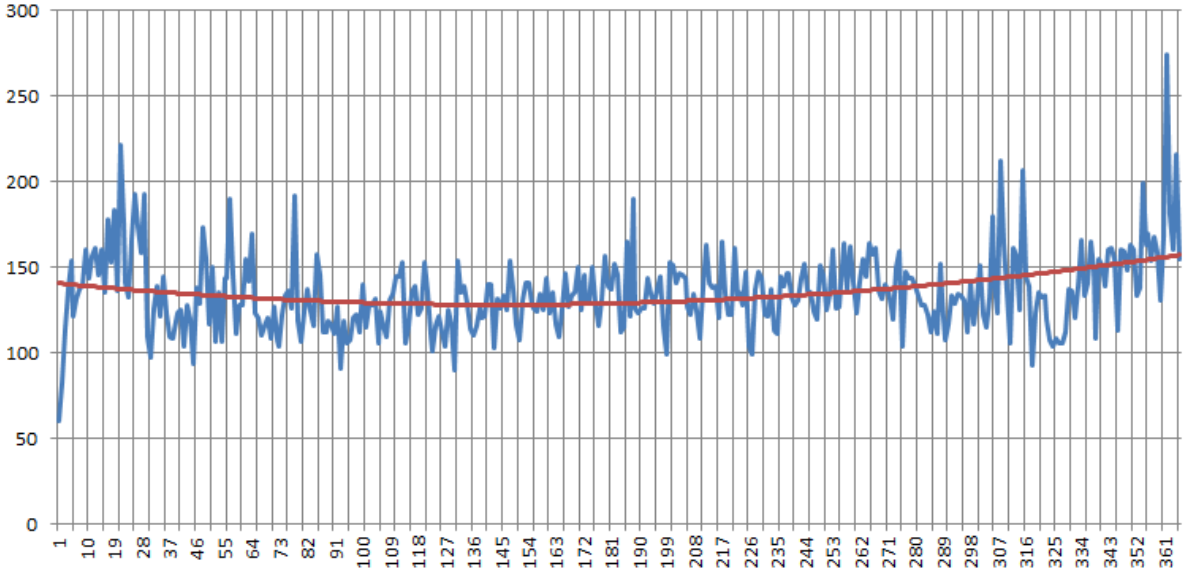

Рис. 4. Изображение тренда на фоне графика временного ряда

Для оценки математической точности полученного уравнения тренда воспользуемся средней относительной ошибкой аппроксимации, которую определим по формуле:

$$A = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \cdot 100\% \tag{16}$$

Получим следующее: A=12,18 %. Таким образом, мы имеем хорошую точность уравнения. Теперь проведем проверку статистической значимости уравнения тренда в целом, используя F-критерий Фишера. Расчетное выражение представлено ниже:

$$F_{расч} = \frac{\sum_{t=1}^{n}\left(\hat{y}_t - \bar{y}_t\right)^2}{\sum_{t=1}^{n}\left(\hat{y}_t - y_t\right)^2} \cdot \frac{n-m-1}{m} \tag{17}$$

где $m+1$ – число коэффициентов уравнения тренда.

В результате расчета получили $F_{расч} = 22{,}825$. Далее определим по табличным данным [4] критическое значение распределения Фишера, учитывая уровень значимости $\alpha = 0{,}05$ и число степеней свободы $df_1 = m = 2$ и $df_2 = n - m - 1$.

$$F_{кр} = 3{,}021$$

Поскольку для нашего случая $F_{расч} = 22{,}825 > 3{,}021$, то с вероятностью ошибки 0,05 уравнение общей тенденции временного ряда в целом статистически значимо (адекватно). Практический интерес также представляет оценка тесноты связи между уровнями ряда и моментами времени. Для этого, используя выражение (18), рассчитаем коэффициент парной корреляции и сравним его со шкалой Чеддока:

$$r_t = \frac{\sum_{t=1}^{n}\left(t - \bar{t}\right)\left(y_t - \bar{y}_t\right)}{\sqrt{\sum_{t=1}^{n}\left(t - \bar{t}\right)^2 \cdot \sum_{t=1}^{n}\left(y_t - \bar{y}_t\right)^2}} \tag{18}$$

Таким образом, получаем $r_t = 0{,}185$, что характеризует практически отсутствие связи. Теперь определим долю вариации y_t от времени t. Для этого вычислим коэффициент детерминации R^2, возведя в квадрат полученный коэффициент парной корреляции.

$$R^2 = 0{,}034$$

Следовательно, количество дорожно-транспортных происшествий за сутки всего на 3,4 % объясняются моментом времени t.

На завершающем этапе уточним нашу модель с учетом сезонности. Как было отмечено ранее, временный ряд обладает недельным циклом на фоне ежедневных данных. Для определения сезонной компоненты s_t нам необходимо найти разности между значениями уровней ряда и их сглаженными значениями $\left(y_t - \tilde{y}_t\right)$, затем отдельно сгруппировать полученные разности по дням недели (понедельник, вторник и т.д.) и найти их средние арифметические. Далее нужно скорректировать полученные сезонные компоненты из условия, что их сумма должна равняться нулю. Для построения модели временного ряда с учетом сезонности прибавим соответствующую откорректированную сезонную компоненту $s_t^{ск}$ к значению тренда:

$$\hat{y}_t^{мод} = \hat{y}_t + s_t^{ск} \tag{19}$$

Результаты расчета сезонной компоненты для наглядности приведены в табличной форме (табл. 3):

Таблица 3.
Значения сезонной компоненты

п/п	День недели	Сезонная компонента $S_t^{ск}$
1	Понедельник	+4,26
2	Вторник	+6,17
3	Среда	–2,95
4	Четверг	+1,17
5	Пятница	+10,76
6	Суббота	–4,12
7	Воскресенье	–15,29
СУММА		**0**

Модельные значения уровней временного ряда, вычисленные по выражению (19) представлены на рис. 5:

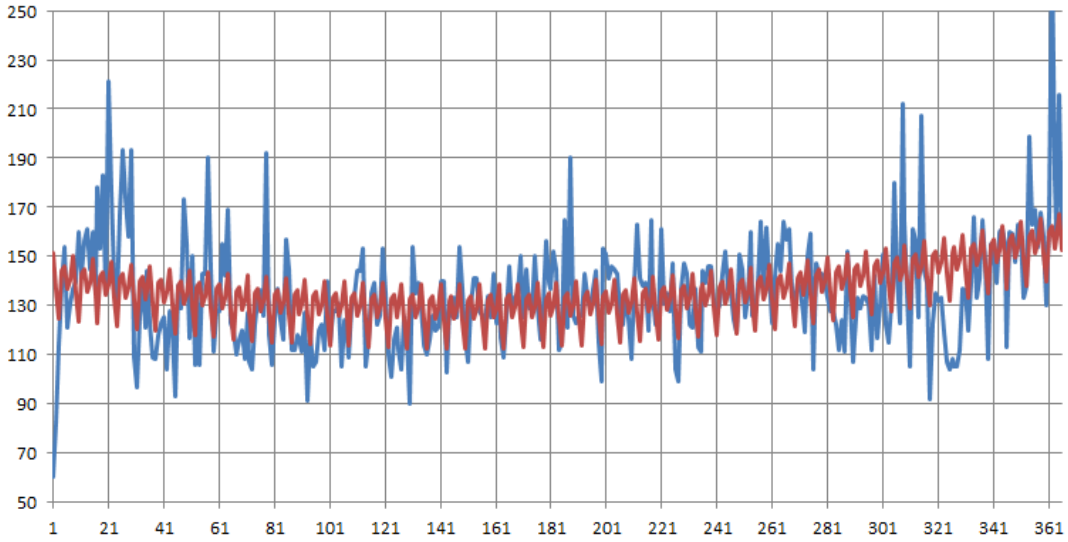

Рис. 5. Аддитивная модель временного ряда с учетом периодической компоненты

Точность полученной модели определим, как и ранее, с помощью средней относительной ошибкой аппроксимации. Величина ошибки, рассчитанная по формуле (16), составила A = 11,11 %

В заключение сформулируем ряд выводов. Исследования ДТП в Нижегородской области проводятся в течение многих лет [1,6]. Обработка статистического материала, характеризующего дорожную аварийность Нижегородской области в 2016 г., обеспечила формирование временного ряда. Анализ ряда позволил построить аддитивную модель, включающую в себя тренд и сезонную периодическую компоненту. Уравнение тренда представляет собой параболу с постоянными коэффициентами a,b,c имеющими смысл ускорения роста уровней ряда, скорости роста уровней и начальных условий. Временной ряд в своем составе содержит периодическую сезонную составляющую с недельным периодом. Максимальное отклонение в положительную сторону наблюдается в пятницу: +10,76 дорожно-транспортных происшествий, в отрицательную сторону в воскресенье: –15,29 происшествий. Теснота связи между уровнями ряда и моментами времени t практически отсутствует. Количество дорожно-транспортных происшествий за сутки на 3,4 % объясняются моментом времени. Построенная аддитивная модель временного ряда имеет хорошую точность A = 11,11 %, на уровне 5 % ошибки статистически значима в целом.

Библиографический список

1. Черевастов, М.Г. Исследование дорожно-транспортной аварийности, зафиксированной в 2016 году на территории Нижегородского региона [Электронный ресурс] // Транспортные системы. 2019. № 3(13). С.11-18.
2. Официальный сайт ГУ МВД России по Нижегородской области [Электронный ресурс]. Режим доступа: https://52.мвд.рф. (дата обращения: 06.12.2019).
3. Карпенко, Н.В. Эконометрика. Анализ и прогнозирование временного ряда: уч. пос. [Текст] / Н.В. Карпенко. – М.: РУТ (МИИТ), 2018. – 132 с.
4. Корн, Г. Справочник по математике для научных работников и инженеров [Текст] / Г. Корн, Т. Корн. – М.: Наука, 1968. – 720 с.
5. Уорсинг, А. Методы обработки экспериментальных данных [Текст] / А. Уорсинг, Д. Геффнер. – М.: Иностранная литература, 1953. – 347 с.
6. Bagichev, S.A**.** Road accident analysis in the Nizhny Novgorod (Russia) and estimation of safety improving facilities [Текст] // S.A. Bagichev, S.R. Biktashev, A.A. Bogdan, S.Yu. Kosin, Yu.I. Molev // FISITA 2010 WORLD AUTOMOTIVE CONGRESS Book of abstracts. Budapest, 30 мая – 04 июня 2010. P. 307.