Санкт-Петербургский государственный университет

Кафедра компьютерного моделирования и многопроцессорных систем

Цаплина Дарья Дмитриевна

Выпускная квалификационная работа бакалавра

Прогнозирование бюджета медицинских учреждений с применением современных методов анализа данных

Направление 010300

Фундаментальная информатика и информационные технологии

Научный руководитель:

PhD,

доцент

Корхов В. В.

Санкт-Петербург

2017

Содержание

Введение

Система здравоохранения - важный социальный институт, который является совокупностью организаций, ресурсов и учреждений, направленных на оказание медицинской помощи. Такая система основана на трех базовых принципах: лечение заболеваний, поддержание здоровья населения и оказание финансовой поддержки в оплате медицинских услуг. Для качественного выполнения функций системы здравоохранения, необходима слаженная работа всех ее компонентов: финансового отдела, компетентных работников, руководства, а также аппарата всеобщего управления. Основной целью такой системы является повышение качества жизни населения за счет оказания услуг, чутко реагирующих на запросы граждан и справедливых с финансовой точки зрения. Важность слаженной работы этого механизма и всех его структур неоспорима так, как это оказывает влияние на качество оказываемых услуг.

Планирование и прогнозирование бюджета больницы сказывается на ее развитие так, как качество предоставляемых населению услуг коррелирует с возвратом затрат на их оказание. Возмещение медицинским учреждениям затрат на лечение граждан РФ оказывают страховые медицинские организации (СМО).

В настоящее время в больницах идет активное внедрение информационных технологий. Однако система работы финансового отдела учреждения здравоохранения требует оптимизации и уменьшения влияния на точность работы человеческого фактора.

Постановка задачи

Медицинские учреждения Краснодарского края обратились с описанной проблемой в ООО ”Виста”. Основная цель деятельности этой фирмы это разработка комплекса программ для повышения производительности внутри фирм. Их основной продукт это медицинская информационная система. Она разработана для управления медицинскими учреждениями. Система дает также возможность использовать:

- онлайн запись на прием;
- диагностические исследования с электронных приборов;
- перечень реестров бухгалтерского и кадрового учета;

Методы прогнозирования в указанной предметной области требуют перехода от интуитивных методов к формализованным методам. Для решения указанной проблемы необходимы: переход от интуитивных методов прогнозирования к формализованным, и разработка системы построения прогнозов, учитывающих допущенные погрешности и чутко реагирующих на запросы населения, тенденции. Проанализировав предметную область, пообщавшись с экспертами в этой области и разобравшись в имеющейся системе, была четко сформулирована задача:

- требуется исследовать экономические принципы работы системы здравоохранения;
- проанализировать расчетные счета за прошедшие года;
- провести исследование, цель которого выбрать метод построения прогноза на месяц, дающий наименьшую погрешность;
- оптимизировать процесс принятия решений по описанной проблеме;

Для выполнения поставленной задачи, можно выделить основные этапы в процессе разработки:

- анализ исходных данных, представленных в виде резервной копии базы данных в формате sql, исходя из предметной области;
- реализация предварительной обработки данных с помощью инструментов MySQL и интерпретируемого языка python;
- исследование полученной выборки;
- обоснованный выбор математической модели для построения прогноза;
- реализация выбранных методов на языке python;

Критерием успеха в выполнении программы можно считать выполнение всех этапов разработки проекта. Получение требуемой программы, способной в реальном времени строить прогноз на ближайший период, равный месяцу, с погрешностью не более 10%.

Обзор литературы

Источниками, которые описывают прикладную область и механизмы распределения денежных средств внутри неё описывает работа [1]. Нормативно-правовые аспекты и ценообразования взятые за основу исходных данных взяты из муниципальных постановлений [2] и федеральных документов [3].

Этапы планирования работы над подобными задачами и подготовка сырых данных к анализу подробно описывает книга [4].

Процесс исследования временных рядов, методы выявления трендовой и сезонной составляющих, и манипуляции с ними приводятся в работах [5] и [6]. Сам процесс моделирования, описание возможных моделей, их достоинства и недостатки, тонкости применения таких моделей на практике и теоретическое описание процесса прогнозирования содержит работа [7].

Инструменты практической реализации математической модели, их полный функционал, описание практических действий применяемых методов приводится в документации соответствующих библиотек для языка python [8], [9] и [10].

Глава 1. Обзор существующих решений

В начале каждого месяца, который равен одному отчетному периоду, сотрудники учреждения строят прогноз, указывающий на какую сумму они окажут услуги в текущий период. В течение месяца этот прогноз может быть изменен, но все поправки сопровождаются документооборотом и проходят через множество структур до полного их утверждения. Это занимает некоторое время, на практике оказывается, что после первой половины месяца, процедура внесения правок не успевает закончиться, до нового периода. После установления суммы, на которую медицинское учреждение окажет услуги пациентам, эти данные отправляются в СМО, которое компенсирует затраты на лечение.

Определение объемов на будущий месяц оказанной медицинской помощи застрахованным лицам напрямую влияет на дальнейшее финансирование учреждения. В случае превышения суммы, на которую оказали услуги, в сравнении с изначальным прогнозом, СМО не возмещает затраты за допущенную погрешность. В обратном случае, это оказывает влияние на дальнейшее финансирование учреждение и объем бюджета на следующий год.

1.1 Интуитивный подход

В настоящий момент основным методом решения выявленной проблемы является интуитивный подход к построению прогноза. Суть подхода основывается на логическом анализе финансовой ситуации, состоянии здоровья населения и опыте финансового отдела.

Эксперт должен провести самостоятельную работу над оценкой тенденций прогнозируемого объекта, его состояний и вариантов развития, а также погрешностей, допущенных в предшествующие года на основе опыта.

Как результат получается, что точность прогноза зависит от компетентности и внимательности каждого специалиста. Прямая зависимость от человеческого фактора влечет за собой неточные оценки, а вследствие этого появляется большая погрешность при построении прогноза. Стоит отметить отрицательный момент данного подхода в том, что при смене сотрудника, опыт полученный за предыдущие года теряется.

Неточно планирование влечет за собой проблемы в финансировании. Но один из худших исходов, которые могут повлечь за собой ошибки при построении прогноза, это закрытие учреждения.

1.2 Коммерческие предложения

На текущий момент не разработано программы, способной решить поставленную задачу в полной мере. Наиболее популярны и широко распространены программы, которые предсказывают продажи определенного товара, либо планируют стратегию складов предприятий. Есть отдельные инструменты, которые в комплексе могут спрогнозировать бюджет учреждения, но для этого весь процесс должен осуществлять компетентный специалист.

На рынке существуют коммерческие продукты предоставляющие инструменты для анализа данных и прогнозирования, лицензия на которые требует больших затрат. К тому же дорогостоящие программы с нужным функционалом не адаптированы под российскую систему финансирования больниц. Также схожим функционалом обладает Excel, но и он требует специалиста, обладающего фундаментальными знаниями математической статистики.

Глава 2. Анализ данных

В качестве исходных данных была предоставлена массивная база данных медицинской информационной системы, которую используют медицинские учреждения на территории Краснодарского края с 2013 года. Предоставленные данные оказались избыточны так, как база данных содержала сотни таблиц с информацией о пациентах, работниках, технике, ресурсах и справочной информации. Для решения задачи нужно определить принадлежность данных и составить информативную выборку, для оптимизации процесса принятия решений.

Проанализировав предметную область, можно сделать формальный вывод, что для решения поставленной задачи необходимы уникальные знания о каждом визите застрахованного лица в учреждение и издержки предоставленных услуг. Для их идентификации и избежания дублирования в указанной системе выборка определяется по уникальным номеру события. А так же затраты на каждую предоставленную услугу, можно вычислить основываясь на таблице содержащей перечень услуг и их стоимость, описывающийся тарифным соглашением в сфере обязательных медицинского страхования по Краснодарскому краю.

Всю вышеперечисленную информацию нужно агрегировать по дням, после этого сформировать выборку и записать в csv файл. Для дальнейшего анализа необходимо посмотреть на получившиеся данные. Для этого построим график, где ось x задают дни недели, ось y сумму, на которую пациентам оказали услуги в данный день.

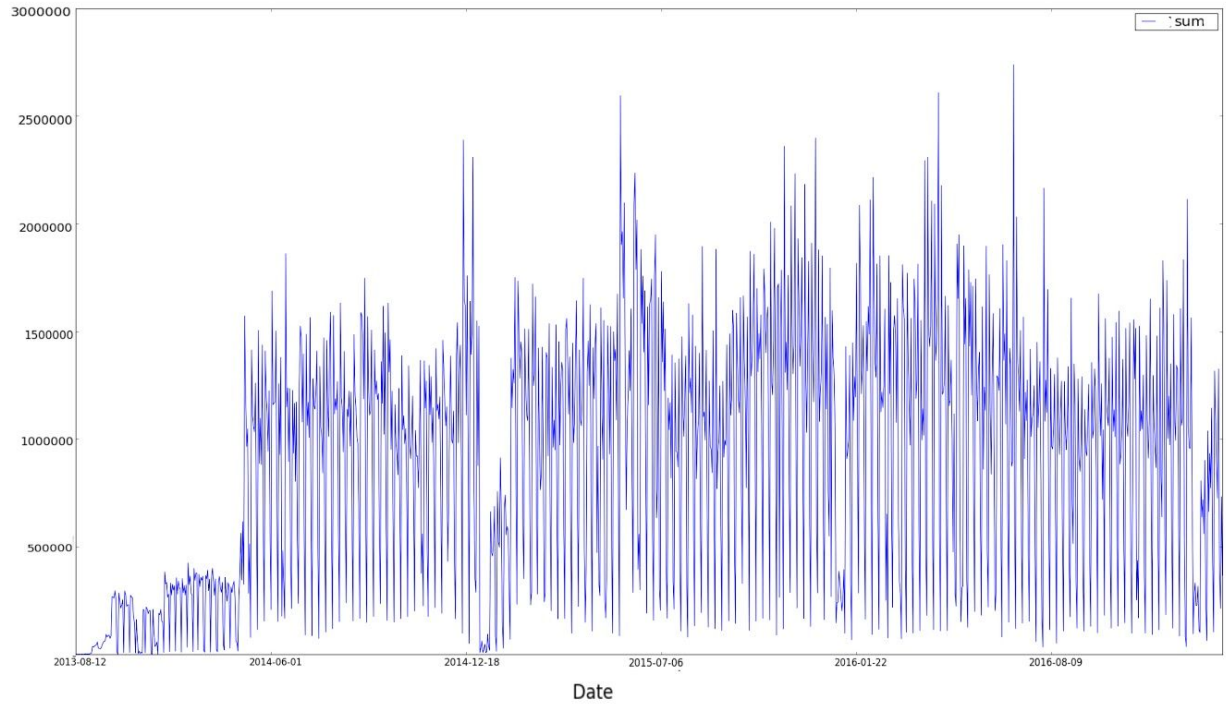

рис 1. Исходные данные

2.1 Подготовка данных

Как видно из рисунка 1, полученные данные оказались грязными, то есть обладающими низким качеством. Процесс предварительной обработки данных до начала анализа необходим для приведения их к соответствующим требованиям, задаваемым предметной областью. В указанной процедуре можно выделить два этапа: очистку, необходимую для повышения качества данных, и оптимизацию, выявление и исключение незначащих признаков.

2.1.1 Очистка данных

Очистка направлена на устранение ошибок в данных с тем, чтобы эти данные адекватно и последовательно представляли процесс, в результате которого они были получены.

Как видно из рисунка 1, данные требовали значительной очистки т.к. оказались неполными и с большими выбросами, которые объясняются

человеческим фактором и достаточно тяжелым процессом внедрения новых технологий в больницы.

Первым этапом очистки данных является устранение противоречий и дубликатов. С первым в решаемой задаче нет нареканий. Все данные логичны и не противоречат друг другу и на уровне описательной характеристики показателей, и на уровне представления их в базе данных. А проблема с дубликатами решаются грамотно разработанной системой, не допускающей повторение данных с помощью уникального номера присваиваемого каждому визиту и пациенту.

Следующим этапом в процессе подготовки данных будет восстановление целостности. Специалисты предметной области не смогли объяснить почему в данных имеются пропуски, то есть даты в которые полностью отсутствует информация о посещениях. Очевидно это является недостоверной информацией, т.к. с минимальной вероятностью возможно, чтобы во всей области не было ни одного обратившегося человека. Поэтому целостность данных была восстановлена путем добавления недостающих дат, а на месте значений записаны нули. Такой подход не скажется отрицательно на результате, обоснование этому приводиться дальше.

В результате качество данных было повышено, что приведет к эффективной работе модели и достоверности результатов анализа. Но данные по-прежнему являются избыточными и обладающими выбросами, поэтому следует провести оптимизацию данных.

2.1.2 Оптимизация данных

Оптимизация данных необходима для избавления от несущественных значений и адекватной работы модели. Следствием неудачной оптимизации

могут служить неточные результаты предсказания в дальнейшем, невозможность извлечь нужные выборки.

Для начала следует объяснить маленькие значения относительно всей выборки в начале 2013 года. Объяснение этому дали специалисты занимающиеся медицинской информационной системой. Процесс перехода на новый вид отчетности и ведения базы данных происходил постепенно, что повлекло за собой увеличение числа учреждений, а значит и числа пациентов. Из этого следует, что для правдоподобности выборки, следует рассматривать более стабильный участок. Поэтому будем вести отсчет с начала 2014 года.

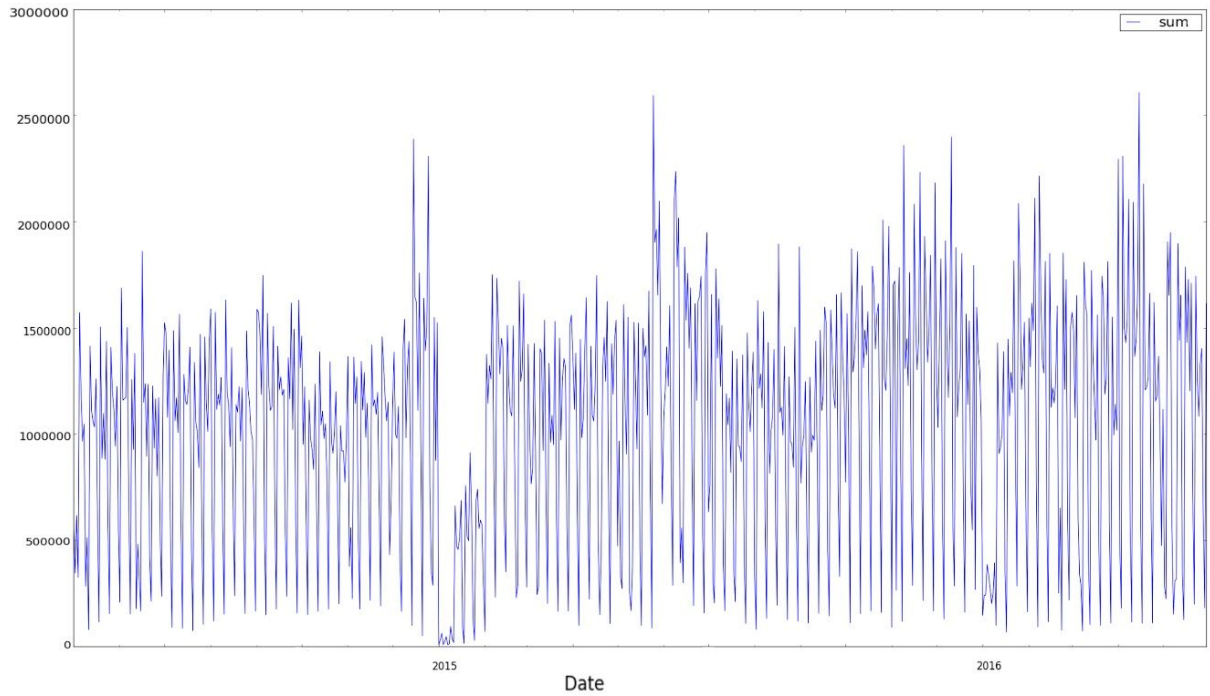

рис. 2 Стабильная выборка данных

2.2 Исследование данных

Перед тем, как приступить к выбору метода прогнозирования, необходимо оценить, что из себя представляют данные. Количественное прогнозирование может применяться, когда выполнены два условия:

1. Имеется информация о прошлом
2. Некоторые аспекты прошлых результатов будут продолжаться и в будущем

Существует широкий диапазон количественных методов прогнозирования, часто разрабатываемых в конкретных дисциплинах для конкретных целей. Каждый метод имеет свои свойства, точность и затраты, которые необходимо учитывать при выборе конкретного метода. В большинстве проблем количественного прогнозирования используются либо данные временных рядов (собираемые через регулярные интервалы времени), либо данные поперечного сечения (собранные в один момент времени).

С данными поперечного сечения мы хотим предсказать ценность чего-то, чего мы не наблюдали, используя информацию о случаях, которые мы наблюдали. Данный подход к прогнозированию не подходит для рассматриваемой задаче.

Данные временных рядов полезны, когда предсказываются наблюдения, которые со временем меняются, например: ежедневные цены на акции Mail.ru, ежемесячные осадки, квартальные результаты продаж для Avito, годовая прибыль Яндекс. Все, что наблюдается последовательно во времени, представляет собой временной ряд.

При прогнозировании данных временных рядов, производится оценка, как последовательность наблюдений будет продолжаться в дальнейшем. Такой ряд состоит из двух элементов: отметках во времени и замерах (значениях), соответствующим указанной отметки времени. С полной уверенностью можно сказать, что мы работаем с временным рядом.

Главной характеристикой временных рядов является их стационарность. Именно эта характеристика в дальнейшем определяет методы, которые применимы к задаче, и этапы дальнейшей работы. Из рисунка 2 видно, что во временном ряде присутствует шум, перед дальнейшим исследованием на стационарность его следует сгладить.

2.2.1 Сглаживание временного ряда

Как видно из графика он обладает большими выбросами, что неблагоприятно скажется на адекватности работы модели. Более детально изучив спады, исходя из предметной области, их можно объяснить человеческим фактором. А именно, каждый спад совпадает с выходными, соответственно учреждения здравоохранения не ведут общий прием в эти дни, а также резкие подъемы выпадают на понедельники, из чего следует предположение, что работники не своевременно заносят данные в информационную систему. Так как повлиять на своевременность ввода данных невозможно, их необходимо сгладить.

Методы сглаживания помогают выделить тренд - повторяющуюся часть временного ряда. При применении метода следует заранее вычислить период, если он существует. В конкретной задаче, вследствие того, что информационная система заполнялась несвоевременно, и существует необходимость перераспределения значений, был выбран метод экспоненциального сглаживания.

Для сглаживания ряда с помощью выбранного метода необходимо рассчитать экспоненциальные скользящие средние.Идея метода заключается в том, что экспоненциальная средняя рассматривается как асимметричная взвешенная скользящая средняя, в которой предшествующие данные берутся

с разными коэффициентами, значения весов коэффициентов убывают по экспоненте в зависимости от удаления от текущей точки.

Пусть $Y = \{y_1, y_2, ..., y_m\}$ есть временной ряд, тогда суть метода описывается в виде рекуррентного выражения:

$$X_t = \alpha \cdot y_t + (1 - \alpha) \cdot X_{t-1},$$

где: X_t - сглаженный ряд, y_t - значение в момент t, α - есть коэффициент сглаживания.

От определения значения коэффициента α зависит насколько предшествующие данные будут оказывать влияние на текущее. Для выбора наиболее оптимального метода вычисления значения α нет. Но есть общий принцип, если начальные условия являются достоверными, следует минимизировать значение α, если же есть сомнения в их достоверности, то следует выбирать большую величину α, что приведет к большему влиянию последних значений на сглаживаемую точку.

2.2.1.1 Реализация метода экспоненциального сглаживания

На практике при выборе малого α, дисперсия в большей степени сокращается, тем самым подавляя колебания изначального ряда, а в случае больших значений α разброс незначительно отличается от дисперсии ряда Y .

Для реализации метода экспоненциального сглаживания, в языке python существует функция DataFrame.ewm(). Один из параметров функции задает значение коэффициента α. Для выбора коэффициента стоит исходить из предметной области, т.к. на рис.1 можно заметить повторяющийся период равный неделе, то предположим, что α=7. Применив сглаживание с указанным параметром, получаем более гладкий ряд относительно исходного:

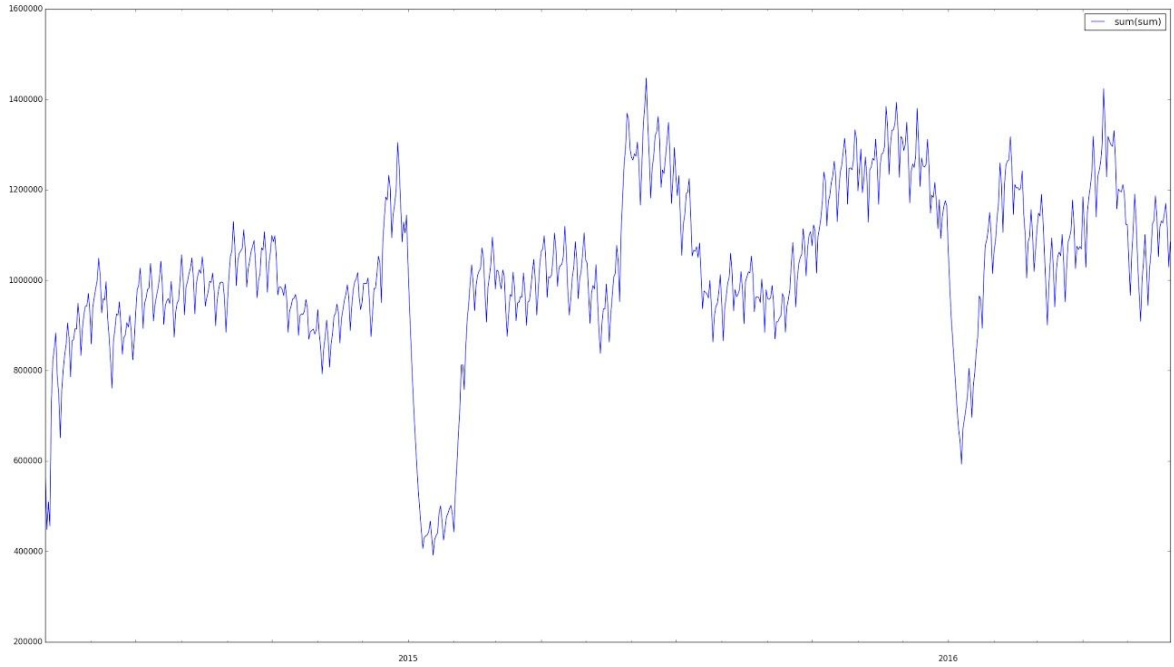

рис. 3 Сглаженный ряд

Можно заметить, что новый временной ряд, при выбранном коэффициенте, не имеет больших выбросов и возможно имеет тренд. Так как спады в начале каждого года можно объяснить новогодними праздниками. Исходя из формального определения временного ряда можно сделать предположение, что ряд нестационарный. Для проверки этого утверждения построим на гистограмму и рассмотрим характеристики ряда.

2.2.2. Описательная статистика

Для оценки однородности временного ряда и характера разброса его значений, стоит прибегнуть к описательной статистики. Временной ряд необходимо охарактеризовать с точки зрения статистики удобно интерпретируемыми показателями. К таким показателям можно отнести: число элементов выборки, среднее значение, минимальное и максимальное значения, стандартное отклонение, квартили. С помощью показателей можно оценить характер выбросов. Построим для этого гистограмму распределения значений:

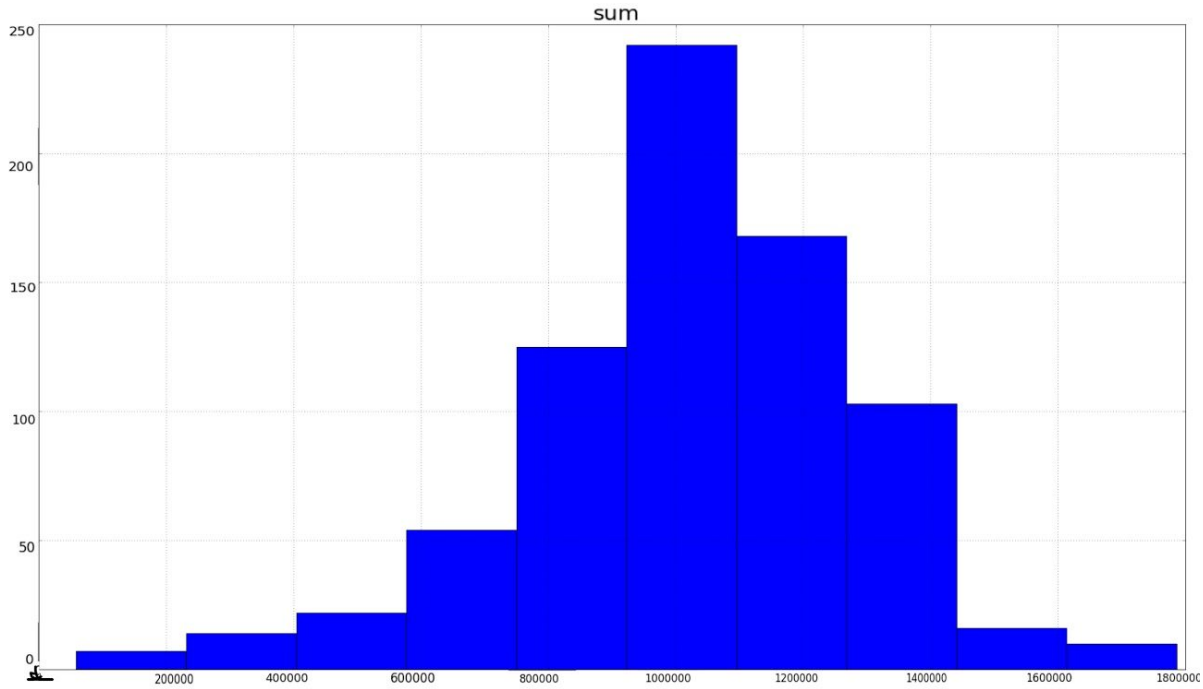

рис.4 Гистограмма распределения значений

Как можно заметить из гистограммы на рисунке 4, получившийся ряд имеет относительно исходного небольшой разброс и однородность. Для более точной оценки следует вычислить коэффициент вариации, который в процентном соотношении относительно среднего показывает однородность разброса значений.

Пусть σ - среднеквадратическое отклонение и $\bar{x}$ - среднее арифметическое выборки, вычисляются по формулам:

$$\sigma = \sqrt{\frac{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}{n}}, \qquad \bar{x} = \frac{\Sigma_{i=1}^{n} x_i}{n},$$

где x_i - значение статистического ряда, n - количество значений в ряду. Тогда коэффициент вариации V вычисляется по формуле:

$$V = \frac{\sigma}{\bar{x}}$$

Если в результате коэффициент получается равным 0%, то ряд является абсолютно однородным. Если он будет больше 33%, то относительно

среднего значения происходит большой разброс, и ряд является неоднородным.

2.2.2.1 Реализация анализа на однородность

Для вычисления коэффициента нам необходимы основные стандартные характеристики ряда. В библиотеке pandas существует стандартная функция describe(), которая предоставляет результаты таких показателей как: количество элементов выборки, медиана, среднеквадратическое отклонение, минимум, максимум и процентиль. В итоге, после её применения, ряд имеет следующие характеристики:

Количество элементов:	761
Среднее арифметическое выборки:	1 021508
Среднеквадратическое отклонение:	184700.2
Минимум:	391828.4
Процентиль 25%	941943.7
Процентиль 50%	1014339
Процентиль 75%	1139627
Максимум	1446973

Рассчитав по вышеописанной формуле, мы получаем коэффициент вариации равный 18%, что свидетельствует об однородности исследуемого ряда. В итоге, данные, однородные, не имеют больших выбросов, оптимизированы и очищены, но для выбора модели прогнозирования определим стационарен ли рассматриваемый ряд с помощью теста Дикки-Фуллера.

2.3 Стационарность ряда

Целью моделирования чаще всего является предсказание значений ряда. А если этот ряд принимает непрерывные значения, точечный прогноз

неинформативен, и требуется пересмотреть подход к определению доверительной полосы при прогнозировании. Из-за этого и некоторых сложностей при построении модели вводится понятие стационарного ряда.

О стационарности временного ряда неформально можно говорить исходя из предметной области. Очевидно, что объем денежных средств относительно времени величина, не изменяющая свои статистические характеристики, соответственно мы можем сделать вывод на основе интуитивного подхода, что ряд стационарен.

С другой стороны, используя аппарат математической статистики, стационарность временного ряда можно установить проверив равенство АКФ (попарное сравнение коэффициентов корреляции каждого порядка с помощью теста на равенство корреляции).

Описанный подход называется тестом Дики - Фуллера. Суть заключается в том, что наличие единичного корня берется за основу нулевой гипотезы, то есть нестационарность ряда. Также рассматривается альтернативная гипотеза о стационарности. Математически это можно описать как:

$H_0 : g = 0$ - на окружности лежит какой-либо корень характеристического полинома, ряд нестационарный

$H_1 : g < 0$ - единичного корня нет, ряд стационарный

Для проверки гипотезы воспользуемся средствами языка python.

2.3.1 Реализация теста Дики - Фуллера

В библиотеке statsmodels в языке python есть функция adfuller(), которая принимает на вход ряд, и проверяет гипотезу о наличии единичного корня и альтернативную ей. Функция проводит несколько тестов способных

определить: результирующее значение теста adf; p-value, полученные с помощью аппроксимации поверхности регрессии; количество лагов; число наблюдений, используемых для регрессии и вычисления критических значений; критические значения для тестовой статистики на уровнях 1%, 5%, 10%. Если значение p превышает критический размер, нельзя отклонить гипотезу, что существует единичный корень. Тест показал следующие значения:

adf	~ -3.801
p-value:	~ 0.002
Критические значения 1%	~ -3.439
5%	~ -2.865
10%	~ -2.568

На практике после получения этих данных рассматриваемую гипотезу можно отвергнуть сравнив результирующее и критическое значения. Если adf больше критического значения, то ряд не стационарен и имеет единичные корни. Полученные значения обратны этому, следовательно, в соответствии с тестом Дикки - Фуллера, который считается фундаментальным, единичных корней нет, а значит исследуемый ряд стационарен.

Глава 3. Моделирование данных

В данной главе будут рассмотрены методы моделирования в широком смысле, тренд и сезонная компонента и то, как они влияют на процесс прогнозирования, обоснование и реализация выбранной модели. Приведены практические результаты и этапы разработки.

Разработка математической модели и поиск оптимальных решений представляет собой итерационный процесс, то есть на каждом следующем шаге, возможен вариант возвращения к предыдущему для улучшения точности и эффективности. В связи с этим, есть необходимость сразу приводить результаты полученные в процессе реализации для обоснования дальнейших действий. Поэтому стоит рассмотреть инструменты, с помощью которых производятся вычисления и моделирование.

3.1 Инструменты реализации

Для реализации всех методов был использован язык python так, как он использовался для написания всех систем разработанных заказчиком, и в дальнейшем для упрощения интегрирования разработанной программы. Python - это высокоуровневый интерпретируемый язык, имеющий простой синтаксис и экономящий время разработчика, что ведет к повышению производительности на практике. Он обладает большим функционалом и применим для широкого круга задач. Для математической статистики и моделирования существуют многофункциональные библиотеки включающие в себя множество тестов, моделей, и способные с помощью визуального представления помогать в интерпретации данных.

Моделирования данных и построения графиков использовалась библиотека Matplotlib. Она позволяет получать визуальное представление

данных в различных форматах в печатном виде и в интерактивных средах на разных платформах. С её помощью можно построить графики, гистограммы, спектры мощности, диаграммы ошибок и разброса, используя всего несколько строк кода.

Для анализа и предварительной обработки данных помимо стандартных функций языка python использовалась библиотека statsmodels, предоставляющая классы и функции для оценки различных статистических моделей, а также для проведения тестов и анализа статистических данных. Помимо этого все результаты тестов проверяются с помощью встроенных пакетов, чтобы убедиться в их правильности.

Но самой базовой библиотекой при математическом моделировании является pandas. Это пакет Python, обеспечивающий быстрые, гибкие и выразительные структуры данных, предназначенные для того, чтобы сделать работу с реляционными данными простой и интуитивно понятной. Разработчики позиционируют его как фундаментальный блок высокого уровня для практического анализа данных в Python.

Scikit-learn это еще одна библиотека, но основной её функционал направлен на реализацию алгоритмов машинного обучения. Но также она имеет различные алгоритмы классификации, регрессии и кластеризации, включая векторные вычисления, градиент, k - средних, и интегрирована с численными и научными библиотеками NumPy и SciPy.

В качестве инструментов разработки использовалась сборка Anaconda, которая включает в себя:

- установленный Python 2.7, 3.4, 3.6
- порядка 300 готовых к установке библиотек и около 150 уже предустановленных (включая все вышеописанные библиотеки)

- установленный IDLE Spider 2

3.2 Основные методы прогнозирования

Возможность прогнозирования чего-либо является необходимым широко распространенным инструментом при процессе принятия решений, когда время и деньги непосредственно связаны между собой. При принятии стратегических решений в условиях неопределенности все делают прогнозы, в большинстве случаев выбор будет направлен на ожидание результатов действия или бездействия.

Можно классифицировать основные методы предсказания, которые часто применяются:

- Экспертные методы прогнозирования (Основывается на оценке эксперта, мнение которого в совокупности представляет единую оценку.)
- Статические методы (Предсказание строится на основе основных показателей описательной статистики: математическое ожидание, дисперсия, различные индексов, вариации.)
- Методы логического моделирования (Рассматриваются для долгосрочных прогнозов используя нахождение закономерностей и поиск)
- Экономико-математические методы (Создаются модели изучаемых областей, и предсказывается дальнейшее поведение при некоторых условиях.)
- Фундаментальный анализ (При прогнозировании анализу подвергаются основные финансовые показатели компании, и ее производительность.)

- Технический анализ (Предсказывает как изменятся значения в будущем на основе анализа изменения значений в прошлом)

Подробно изучив варианты построения прогнозов, был выбран технический анализ. Так как в нем большое количество инструментов и методов, основанных на предположении о том, что при анализе временных рядов, выделив одну из его компонент - тренд, можно предсказать следующие значения. Всего во временном ряду при моделировании и анализе рассматриваются четыре компоненты:

- T - тренд (плавное долгосрочное изменение уровня ряда)
- S - сезонность (циклические изменения уровня ряда с постоянным периодом)
- C - цикл (изменения уровня ряда с переменным периодом)
- E - ошибка (непрогнозируемая случайная компонента ряда)

При техническом анализе для дальнейшего моделирования нужно изучить все вышеперечисленные компоненты.

3.2.1 Тренд и сезонность

Наборы данных временных рядов могут содержать сезонную компоненту, иначе сезонные колебания. Это цикл, который повторяется регулярно со временем, например, ежемесячно или ежегодно. Он может снижать качество данных, либо же наоборот, предоставлять новые знания для повышения точности прогноза. То есть удаление сезонного колебания из временного ряда может привести к более четкой взаимосвязи между входными и выходными переменными. В рассматриваемой задаче определение сезонной компоненты может предсказать периоды гриппа, обязательного прохождение мед.осмотра, эпидемии.

Определение периода сезонности субъективно так, как их можно сколько угодно уменьшать и увеличивать. Самый простой подход к его определению - это анализ в различных масштабах и с добавлением линий тренда. Модель сезонности может быть удалена из временных рядов. Этот процесс называется сезонной корректировкой или дезаминированием. Простым способом корректировки сезонного компонента является использование разности. Если он наблюдается на уровне одной недели, то можно удалить его на текущем дне, вычитая значение с прошлой недели.

Временной ряд имеет явную модель сезонности, а также общую тенденцию увеличения. Мы также можем визуализировать наши данные с помощью метода, называемого декомпозицией временных рядов. Как следует из названия, декомпозиция временных рядов позволяет нам разложить наши временные ряды на три отдельных компонента: тренд, сезонность и шум. Библиотека statsmodels предоставляет удобную функцию seasonal_decompose() для выполнения сезонной декомпозиции:

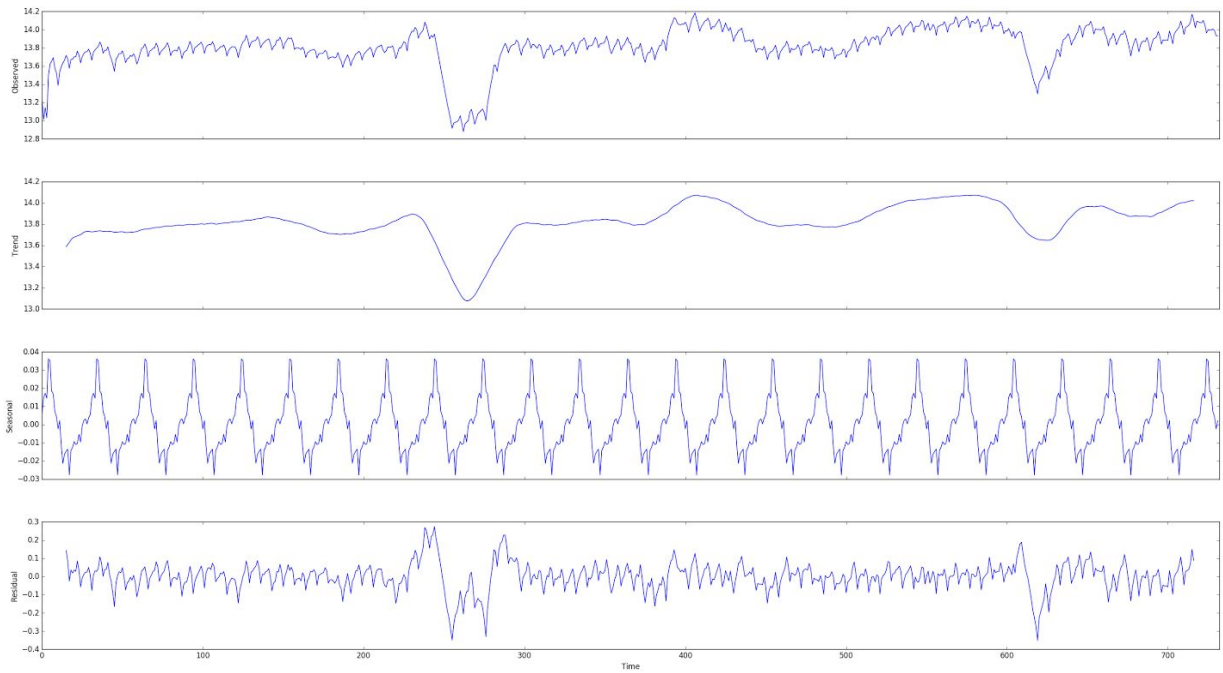

рис. 5 Общий график, тренд, сезонная компонента, шум

На рисунке 5 четко просматривается тенденция к росту наших данных, а также ее сезонная компонента. Они могут быть использованы для понимания структуры исходного временного ряда. Наличие этих компонент важно для разложения временных рядов, поскольку многие методы прогнозирования основываются на этой концепции структурированной декомпозиции для получения прогнозов.

Проводимые тесты показали, что в исходном ряде присутствуют трендовая и сезонная компоненты. Рассматривая предметную область мы также смело можем предположить, что они все же есть так, как есть определенные сезоны, когда прогрессирует грипп и число обращений в медучреждение увеличивается, прохождение медицинского осмотра перед учебным годом учеников и учителей. Также можно сделать предположение, что тренд присутствует и период его равен году так, как вначале каждого года количество обращений в больницы уменьшается. Это подтверждают и спады на графике тренда.

На исходных данных это не особо сказывается из-за достаточно малой выборки, в дальнейшем при накоплении большего объема информации, можно сделать предположение, что эти компоненты будут оказывать большее влияние. В связи с этим стоит рассматривать модели прогнозирования временных рядом способные адаптировать свои методы с учетом тренда и сезонности.

3.2.2 Технический анализ

Модели временных рядов называются иначе стохастическими моделями. На сегодняшний день существует более ста различных методов построения моделей. Рассмотрим наиболее популярные применимые для стационарных временных рядов:

1. Регрессионные модели.

Сюда входят линейная, множественная и нелинейная регрессии. Пусть $w \in W$ - множество параметров, Y - множество зависимых переменных, $x \in X$ - множество свободных переменных. Тогда $f(w, x)$ задает отображение:

$$f : W \times X \to Y$$

Недостатком регрессионного анализа является то, что сложные модели могут переобучаться, а модели, имеющие слишком малую сложность, могут оказаться неточными.

2. Модели экспоненциального сглаживания.

Применяется только для прогнозирования на один период вперед, и наиболее эффективен для среднесрочных прогнозов. Пусть t - период веса, которого мы хотим учитывать; $t + 1$ - период, который нужно предсказать; U_t - экспоненциально взвешенная средняя для ряда; α - параметр сглаживания; Y_t - значение ряда. Тогда прогнозируемый показатель U_{t+1} можно вычислить по формуле:

$$U_{t+1} = \alpha \cdot y_t + (1 - \alpha) \cdot U_t$$

Этот метод нельзя применять для среднесрочного и долгосрочного прогнозирования. Также недостатком является то, что он не учитывает сезонные и случайные колебания.

3. Модели по выборке максимального правдоподобия.

Метод предполагает, что для каждой выборки, предшествующей прогнозу, есть похожая выборка. Она содержится в фактических значениях временного ряда. Функция правдоподобия имеет вид:

$$W(\alpha) = \sum logp(y_i/x_i, \alpha)$$

Для построения прогноза нужно максимизировать значение функции правдоподобия, из множества моделей, то есть выбрать $\alpha^* = argmax_\alpha W(\alpha)$. Недостатком такой модели, является то, что достаточно точные результаты получаются на узком кругу задач, хотя применим этот метод к большинству задач для оценки параметров.

4. Модель на нейронных сетях

Этот подход является громоздким, непрозрачным и достаточно сложным. Чаще всего он используется для распознавания образов, но и в классе задач для прогнозирования показывает высокую эффективность. Недостатком является то, что у разработчика есть возможность формировать входы и наблюдать выходы, но нет возможности проследить как рассчитываются эти значения. То есть нет доступа к тому, что происходит внутри сети. К тому же такие модели требуют большую выборку данных.

5. Авторегрессионные модели прогнозирования

Моделирование основывается на предположении, что последующие или предшествующие значения коррелируют с текущим. Причем прослеживается тенденция к тому, что близко лежащие оказывают большее влияние нежели далеко стоящие, то есть ряд коррелирует сам с собой.

Автокорреляция некоторого порядка и степень связности откликов, разделенных на то же число периодов, относятся к друг другу. То есть механизм прогнозирования основывается на связи между значениями и тому, что она сохранится в дальнейшем.

Пусть n - порядок модели, тогда авторегрессионные модели можно записать как:

$$Y_i = b_0 + b_1 \cdot Y_{i-1} + b_2 \cdot Y_{i-2} + ... + b_n \cdot Y_{i-n} + \varepsilon,$$

где Y - отклик в момент t, Y_{i-n} - отклик на n периодов раньше, $b_1, ..., b_n$ - коэффициенты авторегрессии, ε - случайная компонента.

Недостатком является то, что для построения точной модели необходимо определить ее порядок, а это не так просто. И грань между упрощением модели и точностью прогнозов достаточно сложно определить.

Но все же эти модели показывают высокую точность на среднесрочных прогнозах. К тому же они способны адаптировать свои методы в зависимости от свойств исходного временного ряда. В частности модель ARIMA, при сезонной компоненте равной 0, строит прогноз как ARMA, которая подходит для прогнозирования стационарных рядов без S и T. В связи с тем, что количество данных со временем увеличится, для решения задачи стоит использовать интегрированную модель авторегрессии - скользящего среднего ARIMA.

3.3 Математическая модель ARIMA

ARIMA является интегрированной моделью ARMA, описание ее методов прогнозирование даст понимание о работе выбранного нами метода. Модели ARMA сочетают автокорреляционные методы (AR) и скользящее среднее (MA) в составную модель временных рядов.

Модели MA применяются для обеспечения хорошей адаптации к набору данных, а изменения в этих моделях могут обрабатывать трендовые и сезонные компоненты. Также такие модели используются для создания прогнозов, которые имитируют поведение более ранних периодов. Формулу, где модель основывается на предыдущих данных можно записать как:

$$x_t = \beta_0 x_t + \beta_1 x_{t-1} + ... + \beta_q x_{t-q}$$

$$x_t = \sum_{i=0}^{q} \beta_i x_{t-i}$$

где β_i это веса, для предыдущих значений ряда. Для процесса первого порядка, модель будет выглядеть следующим образом:

$$\widehat{x_t} = \beta_0 x_t + \beta_1 x_{t-1}$$

То есть значение MA оценивается как взвешенное текущего и прошлого значений, такое усреднение является процессом сглаживания независящим от статистической модели. Для того, чтобы охватить процесс скользящих средних в сочетании со случайными процессами, нужно указать статистическую модель. Пусть $\{u_t\}$ - множество независимых, равномерно распределенных случайных величин с дисперсией и нулевым средним. Тогда скользящее среднее порядка q можно записать как:

$$x_t = \sum_{i=0}^{q} \beta_i z_{t-i}$$

Также дисперсия и ковариация x_t вычисляется по формуле соответственно:

$$x_t = \sigma_z^2 \sum_{i=0}^{q} \beta_i^2$$

$$\gamma_t = \sigma_z^2 \sum_{i=0}^{q-k} \beta_i \beta_{i+k} \text{ , k = 0,1,...,q}$$

Используя формулы дисперсии и ковариации можно выразить автокорреляционную функцию (acf):

$$p_k = p(k) = \frac{\sum_{i=0}^{q-k} \beta_i \beta_{i+k}}{\sum_{i=0}^{q} \beta_i^2} \text{ , k = 0,1,...,q}$$

Компонент автокорреляционных методов AR можно выразить как:

$$x_t = \sum_{i=1}^{p} \alpha_i x_{i-1} + u_t$$

где u_t - остаточный член ошибки, α - автокорреляционные коэффициенты. Для процесса AR первого порядка $p = 1$, модель имеет вид:

$$\gamma(k) = \sigma_z^2 \sum_{i=0}^{q} \alpha^i \alpha^{k+i}, \; k > 0$$

Для $|\alpha| < 1$ эта сумма конечна, тогда:

$$\gamma(k) = \frac{\alpha^k \sigma_z^2}{(1-\alpha^2)}, \; k = 0,1,2\ldots$$

$$\rho_k = \frac{\gamma(k)}{\gamma(0)} = \alpha^k$$

Это означает, что для авторегрессионной модели первого порядка автокорреляционная функция определяется последовательно степенями автокорреляции первого порядка с условием . Для стационарных рядов анализ автокорреляции распространяется на модели второго и более порядков. Чтобы адаптировать модель к конкретному ряду нужно стремиться минимизировать ошибку чаще всего с помощью метода наименьших квадратов.

Эти модели можно объединить, просто добавив их вместе как модель порядка (p, q) и записать в виде:

$$x_t = \alpha_1 x_{t-1} + \ldots + \alpha_p x_{t-p} + u_t + \beta_1 u_{t-1} + \ldots + \beta_q u_{t-q}$$

Модель ARMA предполагает, что ряд является стационарным, но чаще всего встречаются случаи, когда существуют тенденции и периодичность, и для её применения необходимо устранить эти эффекты. Удаление заключается в добавлении в модель начальной стадии разности, до тех пор

пока ряд не станет стационарным. Процесс дифференцирования описывается порядком различия.

В совокупности эти элементы описывают тройку (p, q, d), то есть модель ARIMA. После процесса очистки, к полученным временным рядам добавляется модель ARMA. Для выбора параметров следует руководствоваться следующими правилами: модель должна быть как можно более простой, измеренная частичная автокорреляция на лагах 1, 2, 3… должна обеспечить указание порядка AR, форма графика автокорреляционной функции может указывать на тип модели.

Характер разброса	Оптимальная модель
Экспоненциальная, убывающая до 0.	Авторегрессионная модель, следует использовать частичный автокорреляционный график для определения порядка модели.
Чередование положительных и отрицательных значений, спад до 0.	Авторегрессионная модель, следует использовать частичный автокорреляционный график, чтобы помочь определить порядок
Один или несколько выбросов, остальные значения приближены к 0.	Модель скользящего среднего, порядок определяется тем, где график становится равным 0.
Распад, начинающийся после нескольких лагов.	Смешанная модель авторегрессии и скользящего среднего.
Все 0, либо приближаются к 0.	Данные по существу случайны
Высокие значения с фиксированными интервалами	ARIMA, включить сезонный компонент
Отсутствует спад до нуля. Ряд не стационарен	ARIMA, скорее всего ряд не стационарен.

3.3.1 Построение прогноза с помощью ARIMA

Для построения модели будем использовать модель с порядком интегрирования равным 0, так как ряд стационарен, но в дальнейшем при большей выборке данных следует перейти к порядку интегрирования равному 1, то есть строить её для ряда первых разностей. Прежде всего необходимо определить порядок модели. То есть найти оптимальные значения для p и q, компоненты AR и MA соответственно.

Формально поиск параметров можно осуществить по гиперсетке. Взяв матрицу размером 10 на 10, наилучшие значения показали параметры (1;1). Но функции автокорреляции и частичной автокорреляции помогут точно определить параметры.

Для вычисления отличных от нуля корреляционных коэффициентов необходимо построить коррелограмму в MA. Для нахождения максимального ненулевого коэффициента, нужно построить коррелограмму в первой части модели.

Для построения таких графиков библиотека statsmodels имеет стандартные функции. Полученные графики можно интерпретировать следующим образом: Y показывает значения ряда, а X - номера лагов.

Количество лагов в функции определяет число значимых коэффициентов:

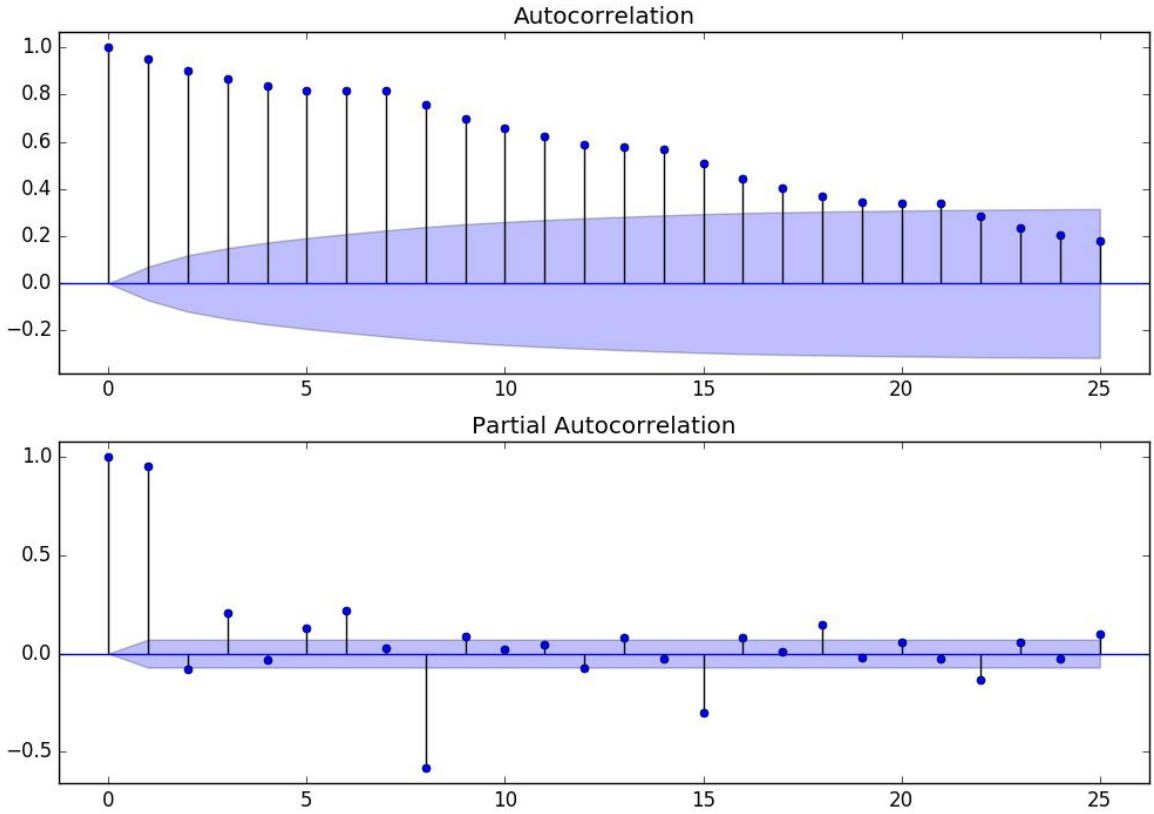

рис. 6 Коррелограммы

По автокорреляции можно наблюдать, что после $x = 1$ значения функции плавно снижаются, но здесь наилучшим вариантом является $q = 1$. Посмотрим на коррелограмму PACF на рисунке 6, на ней наблюдается волновой эффект, начиная со второго значения, наилучший результат при $q = 1$. В итоге, найденые значение совпали с поиском по сетке.

Следующим этапом будет построение модели, но для этого мы возьмем не все данные. Выборку следует разбить на две части: обучающую, на которой модель будет тренироваться, и тестовую, для проверки точности полученной модели. Оптимальным разбиением признано считать 66% на 34%, соответственно. Рассмотрим степень влияния на исходные данные и важность признаков. Для этого построим таблицу коэффициентов:

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6751	0.077	-8.739	0.000	-0.827	-0.524
ma.L1	0.7842	0.062	12.706	0.000	0.663	0.905
ar.S.L12	0.1864	0.097	1.931	0.054	-0.003	0.376
ma.S.L12	-0.5209	0.084	-6.235	0.000	-0.685	-0.357
sigma2	2.737e+09	2e-12	1.37e+21	0.000	2.74e+09	2.74e+09

Важными показателями являются coef (влияние) и P>[z] (значимость), их интерпретация сводится к тому, что все признаки значимые так, как значимость близка к 0.05. Построим диагностику модели, чтобы убедиться в верности предположений с помощью plot_diagnostics():

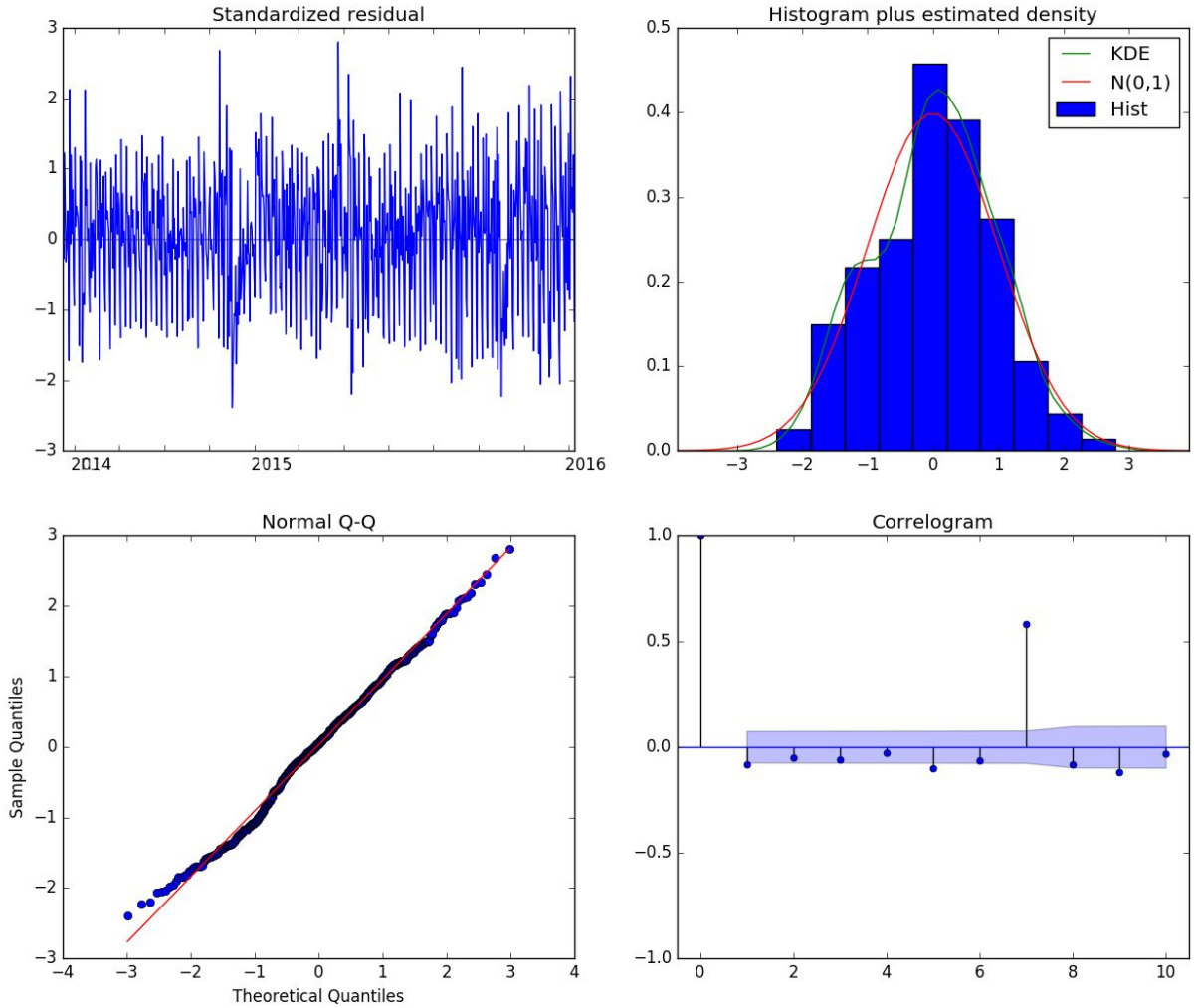

рис. 7 Диагностика модели

Данная модель подходит для построения прогноза, следует это из рисунка 7 верхнего правого графика, где KDE и N(0,1) приближены к друг другу. Значение N(0,1) является нормальным распределением.

Для начала построим прогноз на день, он не является динамическим, а следовательно каждое новое значение предсказывается учитывая всю предшествующую выборку и уже предсказанное значение. Обучив модель с помощью выбранных параметров, получаем прогноз на день:

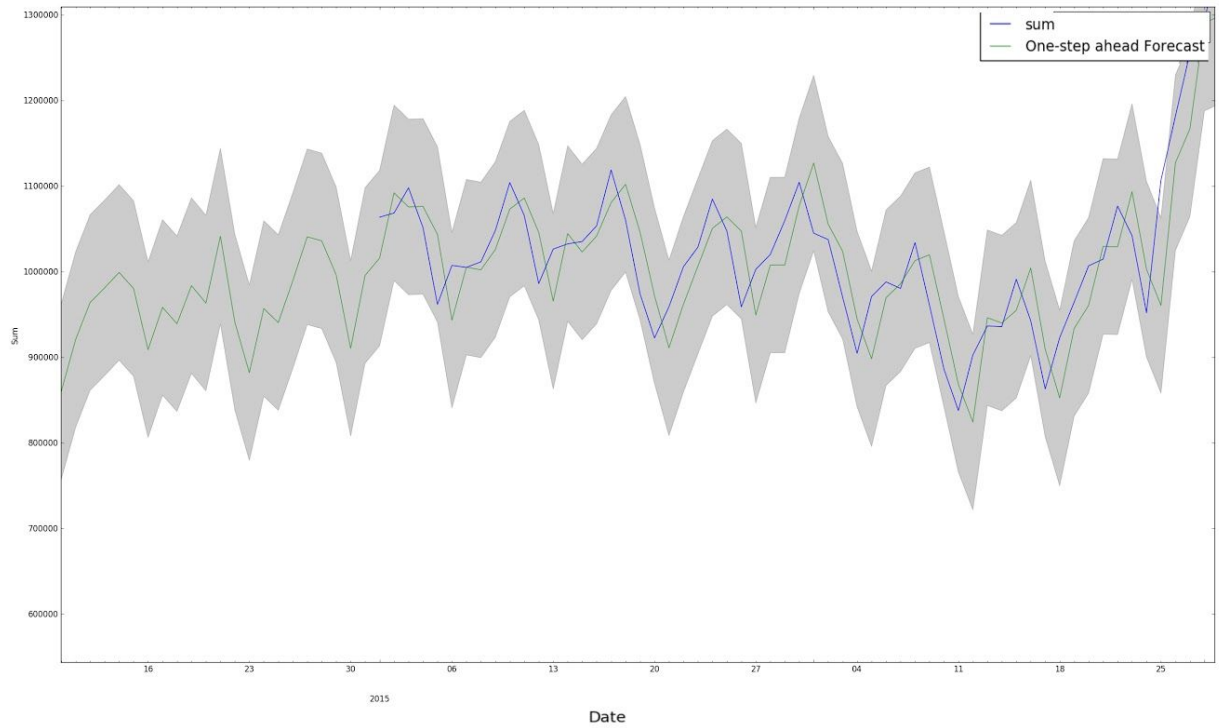

рис. 8 Прогноз на один день

В общем прогнозы показали себя неплохо. Из рисунка 8 видно, что прогноз (синий) лежит внутри доверительного интервала.

Для построения предсказаний на продолжительный период стоит использовать динамическое прогнозирование, оно в отличии от статического, использует значения временного ряда до заданного момента, а после учитывает уже предсказанные значения. Средства используемых библиотек позволяют это сделать, реализованная модель способна строить прогноз на n дней, оно вычисляется как разность между концом отчетного периода и текущим днем. Построим, к примеру, прогноз на 20 дней:

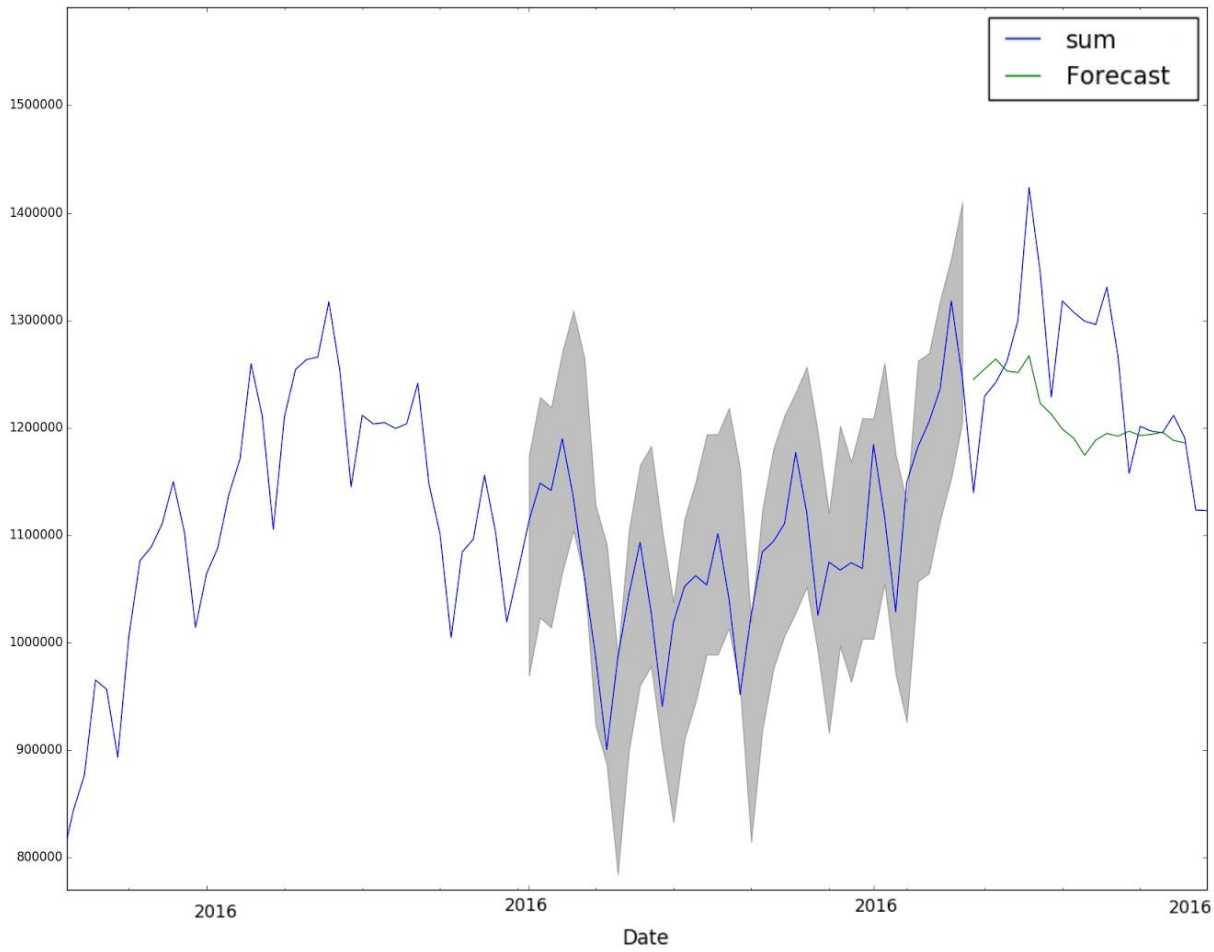

рис.9 Прогноз на 20 дней

Динамическое прогнозирование также выдает хороший результат. Для сравнения на рисунк 9 помимо прогноза (зеленый) выведены данные, которые не использовались в процессе исследования этой работы. Видно, что модель верно предсказала тренд, и достаточно приблизилась к реальным данным, при этом не выходя за границы доверительного интервала.

Выводы

В ходе работы рассмотрены варианты применения математических моделей в системе здравоохранения. Проведен анализ данных, учитывая специфику предметной области. Выбраны оптимальные методы для подготовки данных, их анализа. Рассматриваются математические модели их преимущества и недостатки. Производится обоснованный выбор более эффективной модели, подбор параметров, реализация и построение прогноза.

Оптимальной моделью для текущих данных является ARIMA. Выбраны коэффициенты, при которых модель строит достаточно точный прогноз, при этом улавливая тенденции еще до их появления.

В дальнейшем при увеличении периода наблюдений и минимизации человеческого фактора, учитывая предметную область, можно будет более детально изучить тренд и сезонную компоненту, что должно повысить точность построения прогноза.

Заключение

В результате работы были детально рассмотрены и выполнены поставленные задачи, а именно:

- Рассмотрена предметная область, рассмотрены экономические подходы к решению задачи, проведен анализ данных, выбрана ключевая выборка.
- Выполнена предварительная подготовка данных, путем устранения дубликатов и противоречий, восстановления целостности, оптимизации данных. К полученному ряду применен метод экспоненциального сглаживания.
- Произведено исследование временного ряда на стационарность с помощью различных тестов, в частности Дикки-Фуллера. Проведен анализ трендовой и сезонной компоненты, однородности значений и выбросов.
- Рассмотрены и проанализированы математические модели применимые к рассматриваемой задаче, приведены их недостатки и достоинства. Выбранная математическая модель адаптирована под прикладную область, осуществлен выбор параметров. Приведена реализация и результаты работы прогноза.
- Все рассмотренные методы предварительной обработки, анализа данных, построение математической модели с возможностью прогноза реализованы на языке python с использованием библиотек statsmodels, pandas, scikit-learn. Для визуального представления данных использована библиотека matplotlib.

В качестве результата исследования получен программный инструмент, способный прогнозировать распределение бюджета больницы на

динамический период. И готовый к внедрению в информационную медицинскую систему.

Литература

1. Тарасов Ю. И. Перспективы развития обязательного медицинского страхования // Экономика здравоохранения, 2004. № 3. С. 18-21.
2. Тарифное соглашение в сфере обязательного медицинского страхования по Краснодарскому краю http://www.kubanoms.ru/_files/normativnaya_baza/ts/2017/ts_2017_.pdf
3. Федеральный закон от 21 ноября 2011 г. N 323-ФЗ "Об основах охраны здоровья граждан в Российской Федерации" http://base.garant.ru/12191967/
4. Cielen D., Meysman A., Ali M. Introducing Data Science Big Data, Machine Learning, and more, using Python tools, 2016. 22-48 с.
5. Holt C. C. Forecasting trends and seasonals by exponentially weighted moving averages, 1957.
6. Бокс Дж., Дженкинс Г. Анализ временных рядов: прогноз и управление. Выпуск 1. М.: Мир, 1974. С. 144-164
7. Тихонов Э. Е. Методы прогнозирования в условиях рынка, 2006. М.: Наука. С. 11-49.
8. Documentation Statsmodels python http://www.statsmodels.org/stable/index.html
9. Documentation Pandas python http://pandas.pydata.org/pandas-docs/version/0.15.2/tutorials.html
10. Documentation scikit-learn python http://scikit-learn.org/stable/tutorial/index.html