Лекция 13. Интеллектуальный анализ данных временных рядов в БигДата

Отрасли во всех секторах генерируют и используют данные временных рядов для принятия важных бизнес-решений.

Используя прошлые данные, продуктовая сеть хочет знать, в какое время года рыночный спрос на конкретный продукт достигает пика; колл-центрам необходимо прогнозировать будущие объемы звонков, чтобы иметь достаточный персонал; компании, выпускающие кредитные карты, следят за мошенническими транзакциями — все эти бизнес-решения выигрывают от использования данных временных рядов.

1. Понимание прошлого

Точки данных временного ряда — это снимки прошлого. Понимание исторических событий, закономерностей и тенденций — это некоторые основные показатели, которые отслеживают все компании. Они хотят понять, насколько хорошо они выступали в прошлом и куда они направляются в будущем. В настоящее время, с постоянно растущим объемом «больших данных» и необходимостью получения информации почти в реальном времени, анализ временных рядов становится важной частью бизнес-решений. Вот несколько из многих способов, как исторические данные помогают в принятии решений:

- измерение и отслеживание ключевых показателей эффективности (KPI)
- выявление особых событий, влияющих на спрос на продукцию
- сравнение KPI по категориям продуктов и местоположениям
- понимание тенденции и раннее определение направления бизнеса
- определение сезонности и долгосрочного делового цикла
- гипотетические факторы, влияющие на эффективность бизнеса
- изучение воздействия изменений политики

Базовое понимание исторических событий через данные временных рядов не требует сложного моделирования, просто построение графика зависимости данных от времени может дать очень важные выводы. Раньше электронные таблицы были достаточно хороши, чтобы создавать впечатляющие визуальные истории и идеи. В настоящее время большинство инструментов статистики и анализа данных (например, Python, Tableau, PowerBI) могут достаточно хорошо обрабатывать данные временных рядов для создания диаграмм временных рядов, информационных панелей и т. д.

Исторические данные позволяют получить важную информацию — тенденции, сезонность, деловой цикл и многое другое.

2. Прогнозирование будущего

Еще одно широко используемое использование данных временных рядов — это прогнозирование, то есть составление прогнозов на краткосрочное, среднесрочное и долгосрочное будущее. Многие проблемы прогнозирования, с которыми предприятия обычно сталкиваются каждый день, имеют временное измерение, и именно поэтому данные временных рядов так важны для прогнозирования. Вот несколько приложений для прогнозирования в разных секторах:

- правительства делают прогнозы роста населения для развития адекватной общественной инфраструктуры
- климатологи реконструируют исторические данные и используют их для прогнозов будущего климата

- коммунальные компании прогнозируют будущий спрос для необходимого расширения бизнеса
- предприятия прогнозируют будущий спрос на свою продукцию, чтобы обеспечить достаточный запас продукции

Прогнозирование временных рядов - важная часть технологий науки о данных и машинного обучения, которые включают подгонку статистических моделей / моделей машинного обучения для составления прогнозов.

3. Специализированные бизнес-приложения

3.1 Обнаружение аномалий

Аномалия (или выброс) — это любое необычное наблюдение в данных. Внезапный всплеск транзакций по кредитным картам может квалифицироваться как аномалия, если он является неожиданным с учетом исторического поведения транзакции клиента. Точно так же, если на графиках временных рядов есть пики и падения любых транзакций / точек данных, они также могут квалифицироваться как выбросы.

Обнаружение аномалий — это хорошо изученная область, в которой доступно множество инструментов и методов. Существует несколько пакетов R, предназначенных для обнаружения аномалий, таких как tsoutlier и AnomalyDetection. Вот некоторые бизнес-приложения:

- Финансы: обнаружение мошенничества с кредитными картами и финансовыми транзакциями
- Сетевая безопасность: обнаружение сетевых вторжений на основе аномального трафика в компьютерных сетях.
- Здравоохранение: аномальные состояния пациента на основе считывания записей электрокардиограммы (ЭКГ) и импульсов сердцебиения
- Астрономия: обнаружение выбросов при наблюдении особенностей и характеристик звезд и галактик. Самым известным из всех, вероятно, является обнаружение гравитационных волн (рис. 1).

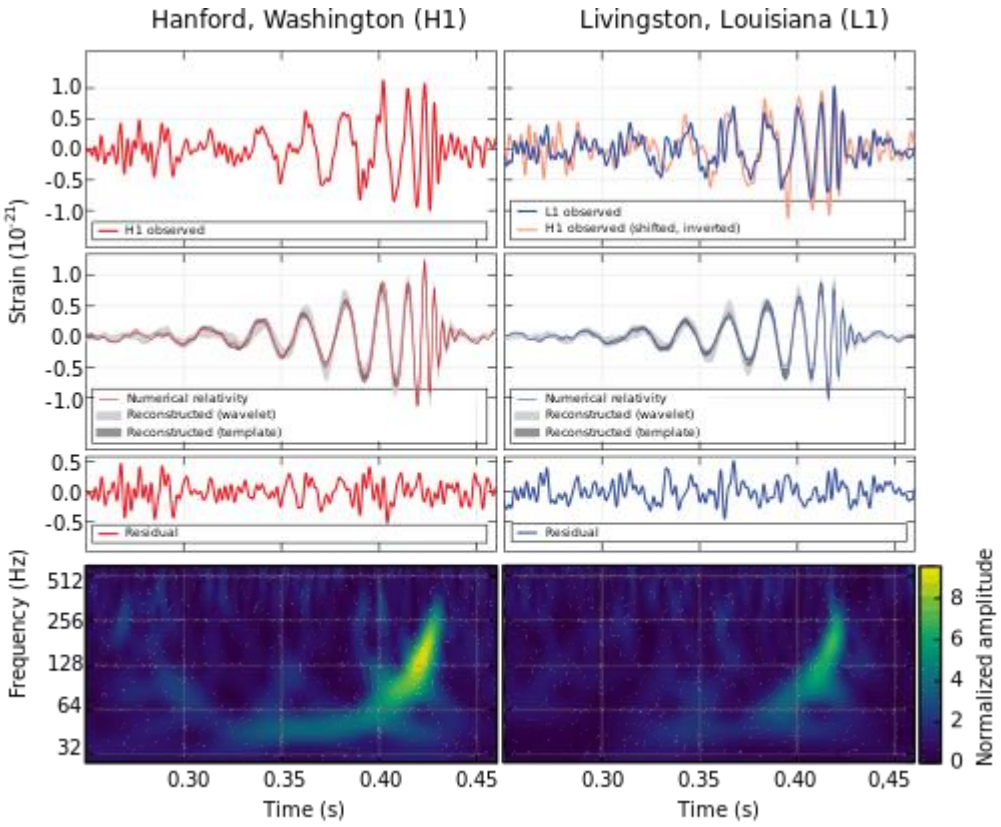

Рисунок 1. Обнаружение гравитационных волн на LIGO: пример обнаружения выбросов

3.2 Сравнительный анализ

Сравнительный анализ — это способ показать сходства и различия между наблюдениями, зависящими от времени. Его также можно использовать для сравнения производительности нескольких объектов. Например, построение временных рядов данных о росте населения по разным странам может дать важную информацию о том, как движется одна страна по сравнению с другой (рис. 2).

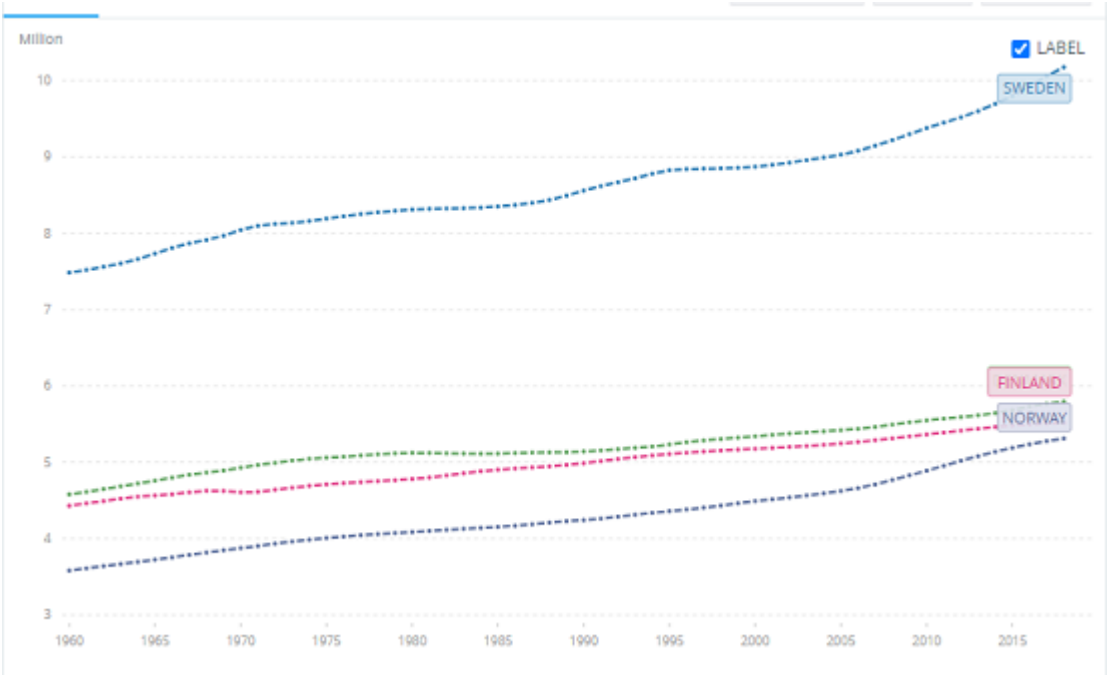

Рисунок 2. График сравнительных временных рядов, показывающий рост населения в странах Северной Европы

3.3 Корреляционный анализ

Корреляция обычно понимается как взаимосвязь между двумя случайными величинами и обычно отображается на двумерных диаграммах рассеяния. Две случайные переменные временного ряда могут точно так же выявить корреляцию / ассоциацию. Построив многомерные временные данные, легко определить связь между двумя характеристиками (например, продажи и прибыль). Эта связь может подразумевать, а может и не подразумевать причинно-следственную связь, но это хорошая отправная точка для поиска характеристик, связанных с выходной переменной, и их использования в статистическом анализе.

Помимо визуализации графиков временных рядов, существуют количественные показатели, которые также используются для измерения корреляций временных рядов - коэффициент Пирсона является одним из таких показателей.

3.4 Эконометрический анализ

Прогнозирующая аналитика - это хорошо известная техническая область, которая в значительной степени опирается на статистику и машинное обучение для прогнозирования будущих результатов. В предметной области также используются методы из специализированной экономической дисциплины - эконометрики . Расширенный статистический анализ, такой как модели панельных данных (например, модели с фиксированными и случайными эффектами), в значительной степени опирается на многомерные продольные наборы данных (временные ряды). Ниже приведены некоторые приложения для эконометрики:

- измерение воздействия изменения климата на ВВП , сельское хозяйство и ущерб от тропических циклонов
- прогнозирование пространственной зависимости между невыполнением обязательств по коммерческой и жилищной ипотеке с использованием данных временных рядов
- использование байесовской точки зрения в задачах прогнозирования временных рядов

- определение детерминант продаж и экспорта на уровне хозяйств

4. Пример использования временных рядов

Компания Nike в 2006 году начала выпускать новую продукцию под названием Nike+, начав с датчика Nike+ iPod для отслеживания данных о физических нагрузках в популярных потребительских устройствах. Эта линейка мониторов активности значительно расширилась и в настоящее время охватывает интеграцию с баскетбольной обувью, беговой обувью, iPhone, Xbox Kinect и носимыми устройствами, включая FuelBand.

Nike+ FuelBand — это носимый на запястье браслет, который содержит ряд датчиков для отслеживания активности пользователя. Мониторинг вашей активности заставляет вас быть более активным, что поможет вам легче достичь ваших фитнес-целей. У Nike есть очень хороший веб-сайт и приложение для iPhone, которые позволяют просматривать графики и детали вашей активности в течение дня, но они не позволяют вам ответить на некоторые более глубокие вопросы, например:

- Есть ли разница в моей активности в выходные и будни?
- Как сезон влияет на уровень моей активности?
- Одинаково ли я активен утром, днем и вечером?
- Сколько я сплю в среднем в день?
- Каково мое среднее время отхода ко сну?

Компания Nike потратила немало времени на разработку программного интерфейса приложений (API) и сопутствующих сервисов, чтобы приложения могли считывать и записывать данные с датчиков, и ниже представлен набор графиков, показывающих информацию об уровне активности одного пользователя по данным Nike+ FuelBand (данные предоставляются каждому пользователю конфиденциально, данный пользователь опубликовал свои результаты).

Сезонный анализ

Используя данные с сентября 2012 года по октябрь 2013 года, он создал набор данных, содержащий около 560 000 наблюдений. Одним из самых простых способов изучения таких больших данных является сезонный анализ. Сезонные данные можно обобщить с помощью сезонного индекса. Например, рисунок 3 получен с помощью почасового сезонного анализа

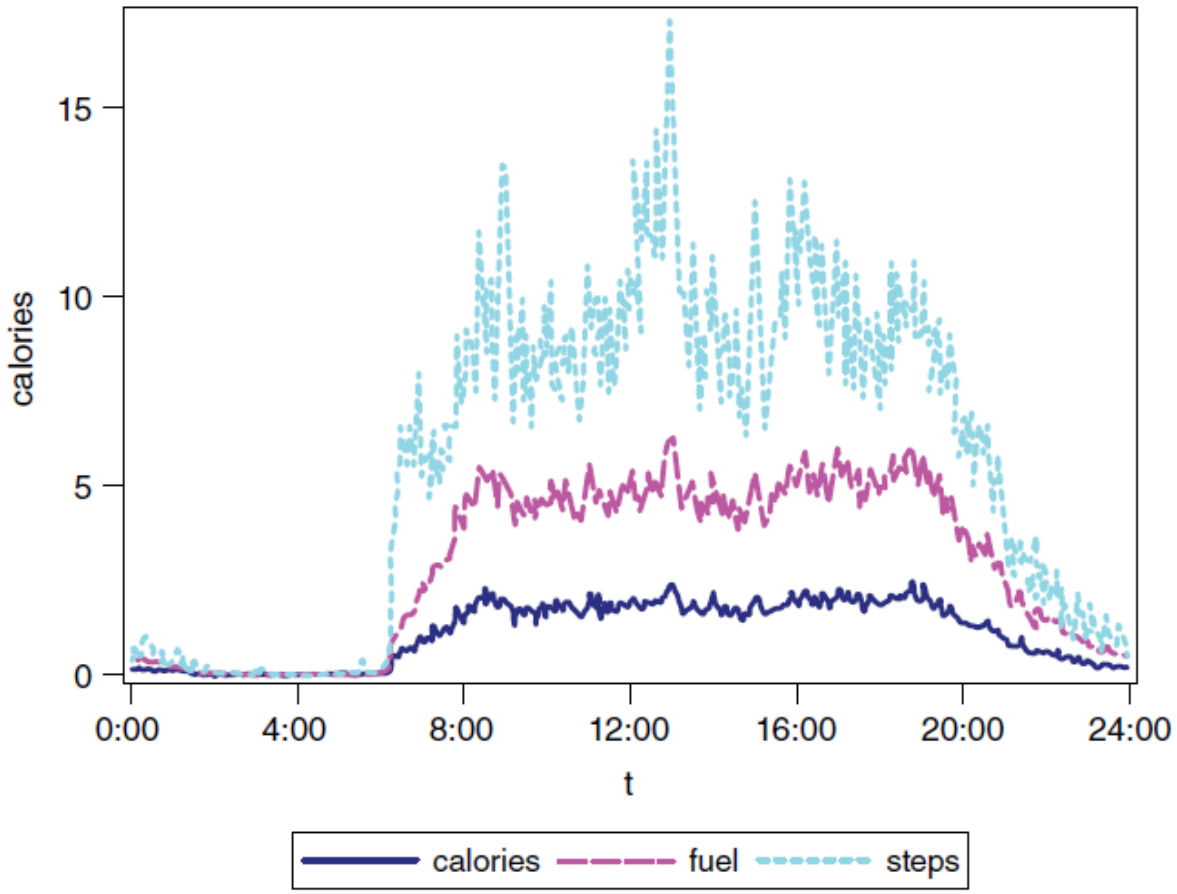

Рисунок 3. Сезонный индекс уровня активности пользователя

Из рисунка 3 видно, что его день довольно рутинный. Как сообщает он сам, «Обычно я становлюсь активным вскоре после 6:00 утра и повышаю уровень своей активности до 9:00 утра, когда я обычно прихожу на работу. Затем я поддерживаю тот же уровень активности в течение всего рабочего дня и до вечера. Около 7:00 вечера я начинаю снижать уровень активности, пока не отхожу ко сну, что происходит примерно с 12:00 до 1:00 ночи. Это очень точная картина моего типичного дня на макроуровне. В будние дни я ставлю будильник на 6:30 утра, а с четырьмя маленькими детьми я встаю в 7:00 утра в большинстве выходных. У меня есть предсказуемая утренняя рутина, которая раньше включала в себя пятимильную пробежку. Рабочий день обычно включает перерыв на обед и короткую прогулку, которая проявляется в виде небольшого всплеска около 13:00. Моя дневная активность снижается в 19:00, потому что в это время мои дети ложатся спать, и уровень моей активности соответственно падает».

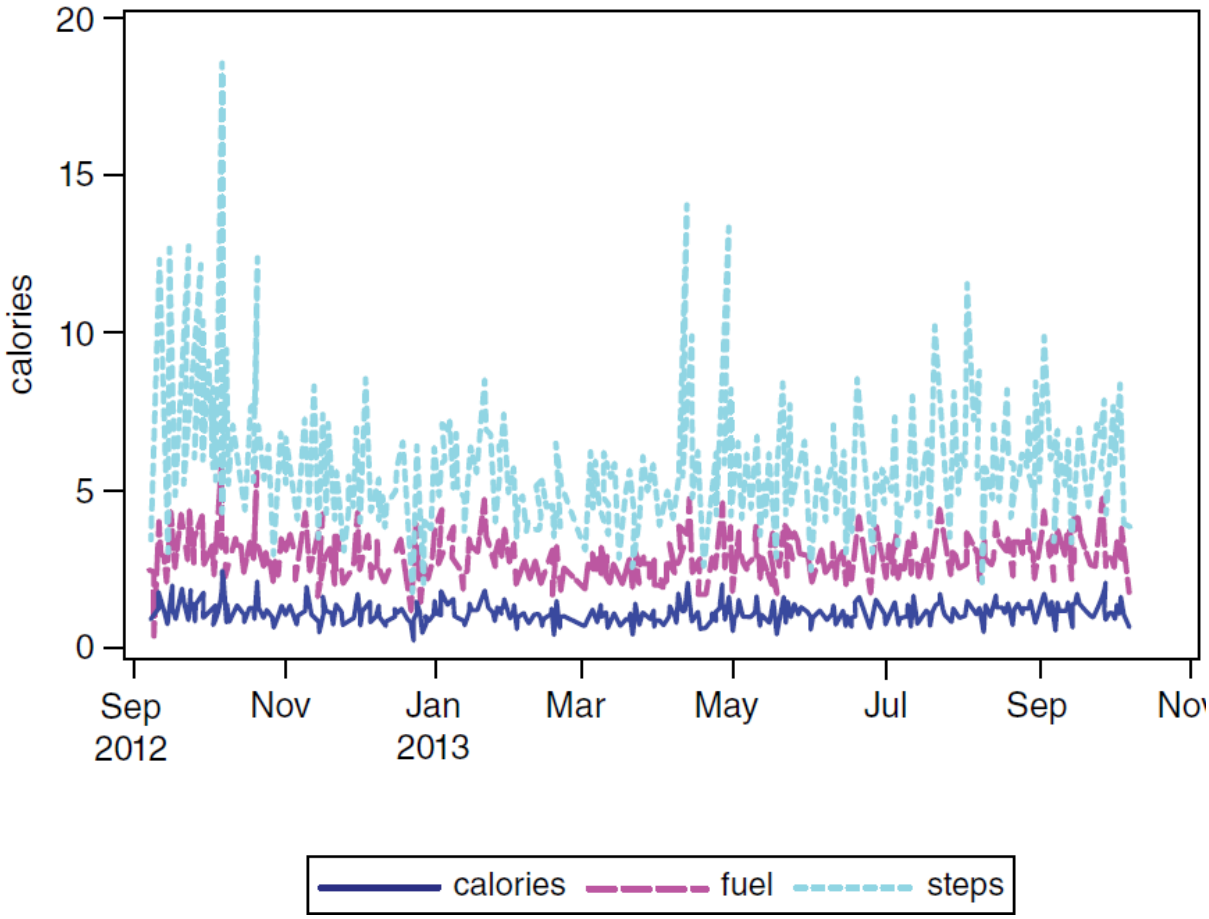

Рисунок 4. Уровень активности от минуты к минуте за 13 месяцев

Анализ тенденций

На рисунке 4 виден уровень активности в виде тенденции за 13 месяцев. По словам хозяина браслета, «Виден первоначальный рост активности, так как мой уровень активности стал азартным благодаря возможности отслеживать его. Также наблюдается типичный всплеск уровня активности, который часто происходит в начале нового года и обязательства быть более активным. Кроме того, в начале мая произошел еще один всплеск, когда погода стала теплеть и мои дети стали участвовать в футбольных, футбольных и других лигах, что заставило меня выйти на поле и быть более активным. Кроме того, дни стали длиннее, что позволило больше времени проводить на улице. Эта тенденция сохраняется на протяжении всего лета».

Анализ сходства

Посмотрим на эти данные с другой стороны. Поскольку FuelBand является непрерывным биосенсором, можно обнаружить аномальный дневной паттерн (если таковой существует). Это и есть проблема поиска сходства между целевыми и входными сериями. Данные были использованы за первые два месяца, чтобы составить запрос (целевой) ряд, который представляет собой почасовую последовательность, усредненную за два месяца. Другими словами, это среднее значение

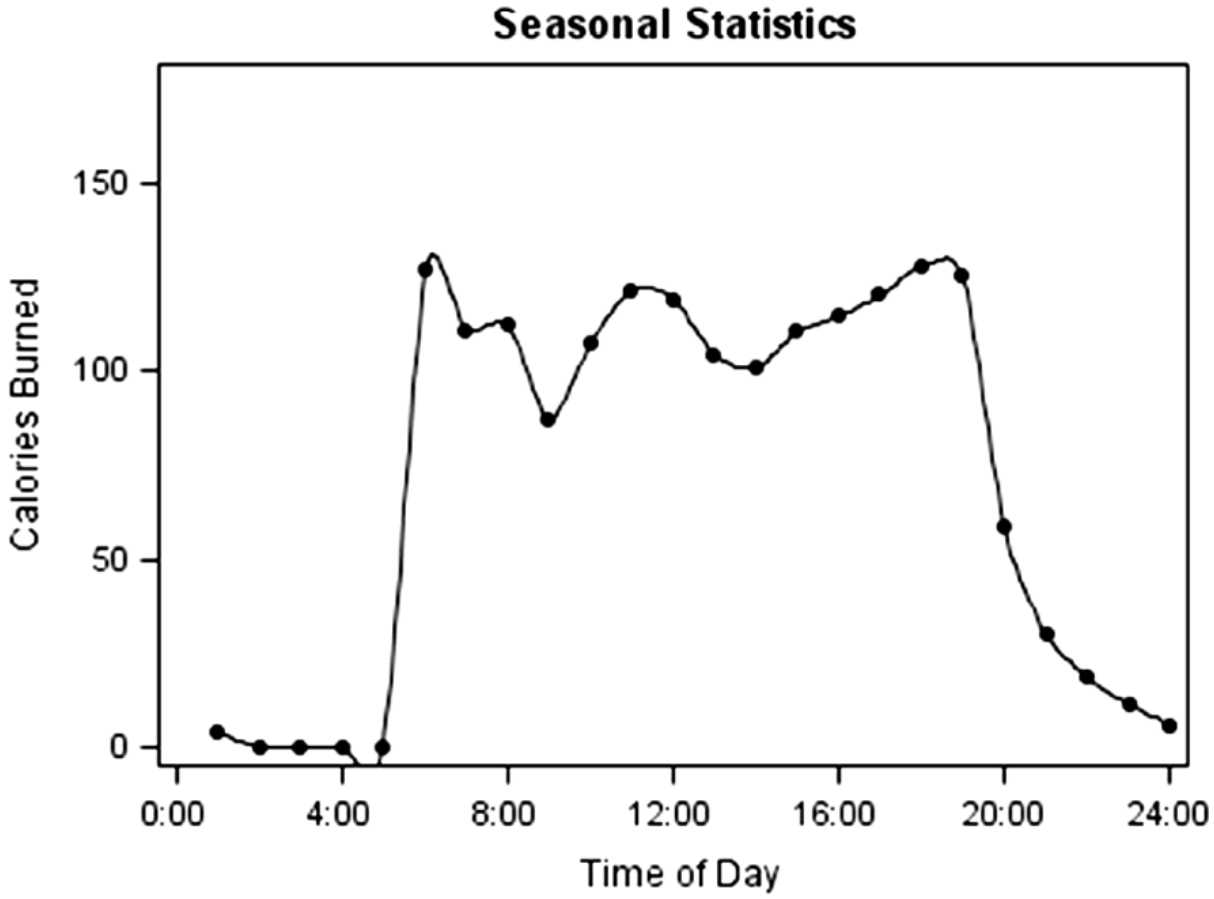

Рисунок 5. Средняя схема сжигания калорий за двухмесячный период в виде почасовой схемы за день.

На рисунке 5 показана усредненная почасовая структура сожженных калорий с 09 сентября 2012 года по 9 ноября 2012 года. Используя процедуру SIMULARITY в SAS/ETS R, был найден аномальный день на основе данных о сожженных калориях. Используя методы динамического искривления времени для измерения расстояния между двумя последовательностями, были найдены пять наиболее аномальных дней на основе исходных моделей за первые два месяца данных по сравнению с примерно 300 ежедневными сериями с 10 ноября 2012 года. Оно имеет множество применений в таких областях, как движение клиентов в розничном магазине или мониторинг сетевого трафика в сети компании.

Из таблицы 1 видно, что 1 мая 2013 года и 29 апреля 2013 года были двумя самыми аномальными днями. Этот показатель ненормальности не говорит о том, что данный пользователь был более или менее активен, скорее о том, что характер его активности отличался от рутины, которую он установил в первые два месяца. По словам пользователя, эти два дня соответствуют конференции, SAS Global Forum. Во время конференции он посещал доклады, встречался с клиентами и демонстрировал новые предложения продуктов в демонстрационной зоне программного обеспечения.

Таблица 1. Пять дней с наибольшим отклонением в суточной активности

Наблюдение	**Последовательность дней**	**Расстояние до базовой серии**
1	ActDate_5_1_2013	30.5282
2	ActDate_4_29_2013	25.4976

3	ActDate_1_13_2013	18.6811
4	ActDate_8_22_2013	17.6915
5	ActDate_6_26_2013	16.6947

Как видно из рисунков 6 и 7, его образ жизни на конференции сильно отличается от обычного распорядка дня. он встает утром позже и в целом менее активен, но уровень активности повышается в конце дня.

Для некоторых видов анализа наибольший интерес представляет аномальное поведение, для других — сходство с закономерностью.

Примером аномальных паттернов является наблюдение за хронически больными или пожилыми пациентами. С помощью этого типа анализа сходства можно удаленно наблюдать за пожилыми людьми, находящимися на дому, а затем направлять медицинский персонал, если поведение отклоняется от нормы.

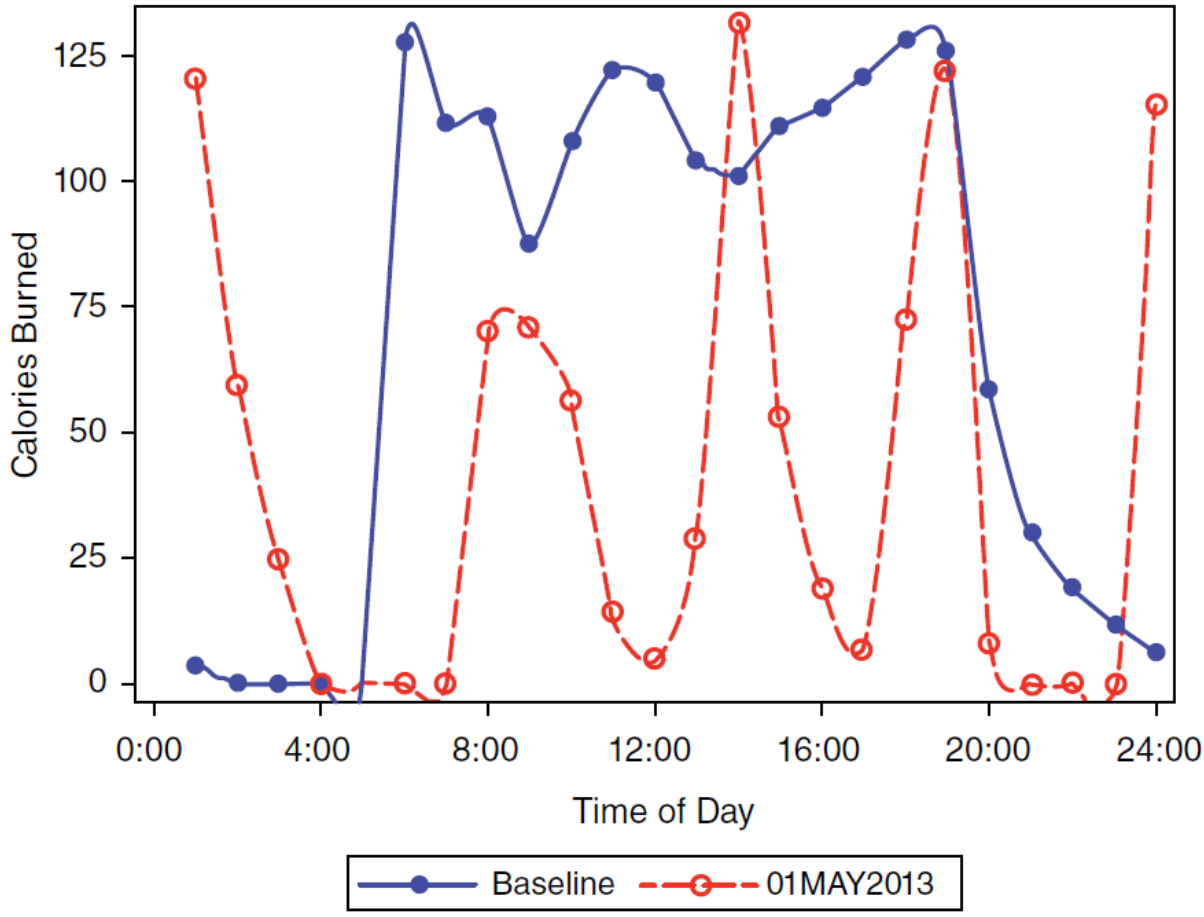

Рисунок 6. Сравнение базовой серии и 01 мая 2013 г.

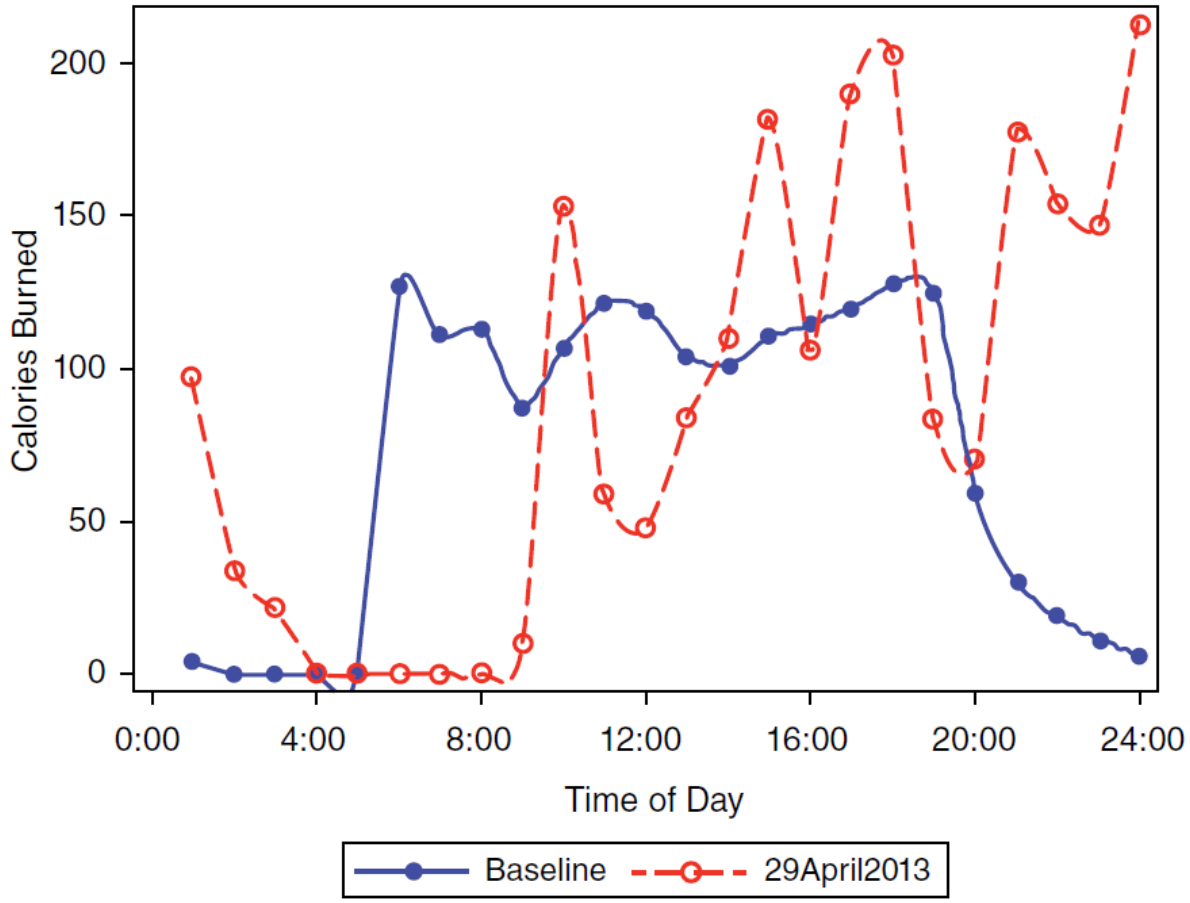

Рисунок 7. Сравнение базовой серии и 29 апреля 2013 г.

Этот метод анализа сходства полезен, поскольку у каждого пациента будет свой уникальный паттерн рутинной активности, который можно сравнить с его текущей дневной деятельностью. Примером сходства паттернов является бег на марафонскую дистанцию. Если вы пытаетесь отобраться на Бостонский марафон, то, просматривая свои индивидуальные километровые отрезки, вы захотите увидеть модель, типичную для других участников[1]. На рисунке 8 показаны два наиболее похожих дня из данных FuelBand.

Сейчас наблюдается сильная тенденция к созданию носимых устройств, которые собирают постоянное количество подробных данных.

[1] Бостонский марафон - единственный марафон в Соединенных Штатах, для участия в котором каждый участник должен пройти квалификационный отбор. Квалификационное время определяется по возрасту и полу.

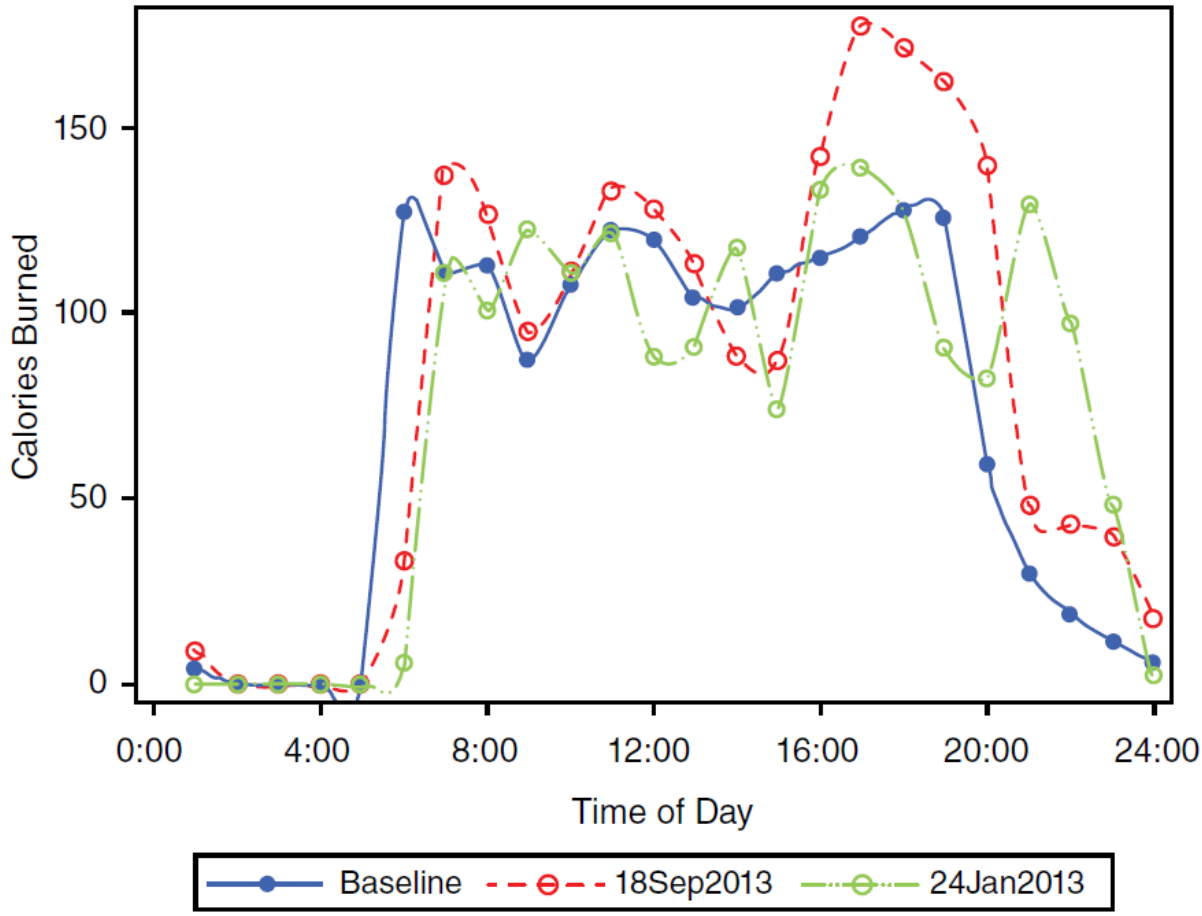

Рисунок 8. Два наиболее похожих дня на базовый уровень скорости каждые 10 секунд во время работы автомобиля

Выводы

Данные временных рядов предоставляют множество возможностей для аналитики и приложений во всех областях приложений. Исторический анализ, прогнозирование, обнаружение аномалий и прогнозная аналитика — лишь некоторые из этих возможностей. Новые аналитические рубежи также появляются с развитием новых инструментов и методов. Искусственные нейронные сети (например, LSTM) и эконометрика являются передовыми рубежами в области анализа данных временных рядов. Как опытные, так и начинающие специалисты по данным могут внести огромный вклад в свои области, воспользовавшись этими инструментами или, возможно, разработав новые.

Список использованных источников:

1. Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners (Wiley and SAS Business Series) 1st Edition.

2. Mahbubul Alam. Time series data mining techniques and applications. URL: https://towardsdatascience.com/time-series-data-mining-techniques-and-applications-5159b8ad26ec (Дата обращения: 22.09.2020)