**Поиск ошибочных выбросов во временных рядах координат ГНСС марок**

# *Аннотация. Применение моделей машинного обучения для выявления выбросов во временном ряде GPS координат марок Центрально-Азиатской сети.*

# *Сравнение моделей экспоненциального взвешивания, линейной, Ridge, Lasso, регрессией моделей решающего леса и градиентного бустинга, XGBoost. Выбор параметров перекрестной проверки с функций потерь для качества и скорости обучения алгоритма.*

# *Очень схематично и кортко, не понятно.*

# *Ключевые слова: Линейная регрессия,* *экспоненциальное сглаживание, временной ряд, перекрестная проверка.*

После обработки GPS данных на выходе программы GAMMIT/GLOBK получают высокоточные временные ряды и на их основе векторы скорости для пунктов наблюдения. Один из последних циклов обработки GPS данных Центрально-Азиатской GPS сети включает в себя информацию по трем координатам для >1600 марок за 1994-2018 годы наблюдений. Для получения корректных выводов по этим данным необходим визуальный анализ временных рядов координат для выявления в них ошибочных аномалий, которые могут возникать на разных этапах формирования данных от процесса измерения и до этапа обработки. Причиной таких аномалий может быть человеческий фактор, аппаратурные сбои, воздействие неизвестных внешних факторов. Во временном ряду координаты это может выражаться в интервалах пропуска данных, ступенчатое смещение тренда и отскоки позиции от линии основного тренда.

Визуальный анализ временных рядов координат каждой GPS марки занимает значительное количество времени, в результате чего возникла необходимость автоматизировать этот процесс. В данной работе интерес представляет задача выявления выбросов во временном ряду и возможная их коррекция с использованием, как самых простых методов, так и методов с машинным обучением. Для анализа нами взят временной ряд центрированных значений восточной компоненты для марки IAT3 которая измеряется непрерывно с 2005 по 2018г (рис.1).

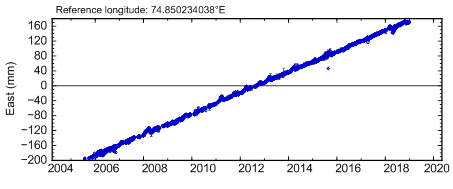
******

Рисунок 1 - Центрированные значения восточной координаты марки IAT3 в системе отсчета ITRF-2014/

Из графика видно, что данный ряд имеет небольшое количество выбросов, которые слабо влияют на разброс и ярко выраженный тренд изменения координаты.

В общем случае при анализе временных рядов выделяются две основные задачи: идентификация и прогноз. Задача идентификации какого-либо события предполагает определение параметров анализируемой системы: размерность вложения, корреляционная размерность, энтропия и др. Задача прогноза имеет цель на основе известных данных наблюдений предсказать будущие их значения, т.е. составить прогноз на некоторый отрезок времени вперед.

В этой работе для нас интерес представляет только задача идентификации выбросов, то для анализа можно применить простые методы скользящего среднего и экспоненциального сглаживания. Метод скользящих средних имеет элемент субъективности и зависит от длины периода L, выбранного для вычисления средних значений. Для того чтобы исключить циклические колебания вдоль генерального прямолинейного тренда, длина периода L должна быть целым числом, кратным средней длине цикла. Этот метод хорошо обозначает тренды, чем больше ширина интервала, тем более сглаженным будет тренд.

Все приведенные в этой работе расчеты будут проводиться на языке Python, и в среде разработке jupyter-notebook, которая работает в браузере. Исходя из поставленных нами задач при анализе данных в была импортирована библиотека Pandas [1].

На рис. 2 отражен график скользящего среднего с ширеной окна 30 дней и доверительным интервалом ±3σ, за границы которого попадают выбросы.

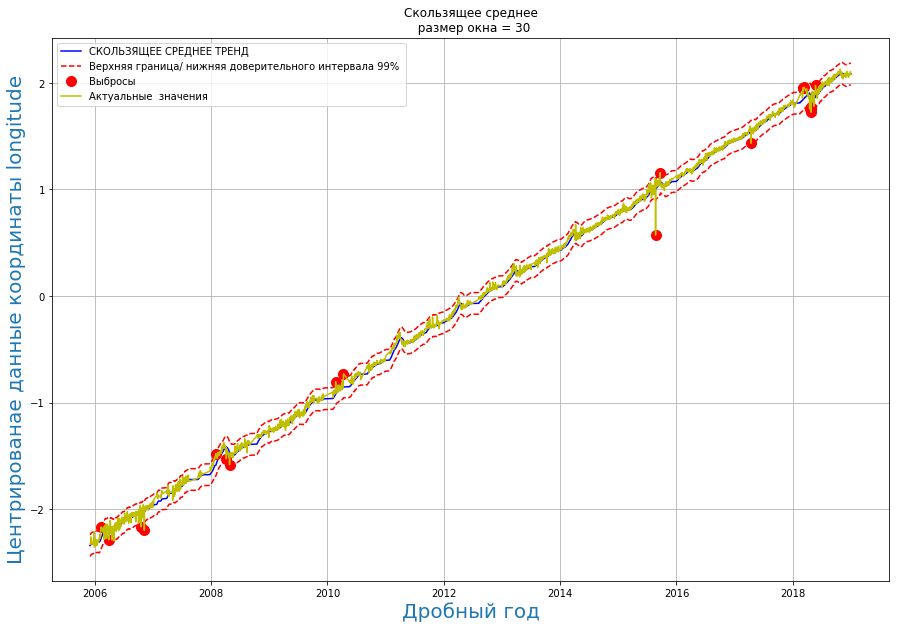


Рисунок 2 – График скользящего среднего с ширеной окна 30 дней и доверительным интервалом ±3σ. Красные точки идентифицированные выбросы по центрированным значениям долготы.

Подпись вертикальной шкалы: (E-Eср)×10-6, °

Выделенные таким образом выбросы превышают коридор случайных отклонений в ±3σ от усредненного тренда и могут быть удалены из временного ряда данных, или уменьшены до допустимых граничных значений в ±3σ.

Модификацией простого скользящего среднего является метод экспоненциального сглаживания, который получил свое название от последовательности экспоненциально взвешенных скользящих средних. Каждое значение в этой последовательности зависит от всех предыдущих наблюдаемых значений. При экспоненциальном сглаживании веса, присвоенные наблюдаемым значениям, убывают со временем, поэтому наиболее часто встречающиеся значения получат наибольший вес, а редкие величины — наименьший. В этом случае используется боле сложная модель экспоненциального сглаживания.Модель Хольта-Уинтерса [2], является расширением метода Хольта до трехпараметрического экспоненциального сглаживания. Это значит, что метод характеризуется тремя параметрами, которые необходимо выбрать, чтобы получить прогноз. Выбор этих параметров может производиться путем простого перебора.

Вес α называется сглаживающим фактором, отвечает за сглаживание ряда вокруг тренда, β за сглаживание самого тренда и третий параметр отвечает за цикличность (γ). Выбор сглаживающего коэффициента, или веса, присвоенного членам ряда, является принципиально важным, поскольку он непосредственно влияет на результат. К сожалению, этот выбор до некоторой степени субъективен, поэтому перебором параметров будет заниматься перекрестная проверка [3] с использованием метрики качества [4].

Процесс кросс-валидации в себя включает: разбиение набора данных на части (фолды), затем построение модели на одной части (называемой тренировочным набором). Чтобы уменьшить разброс результатов, разные циклы кросс-валидации могут проводится на разных объемах и разных комбинациях исследуемого набора данных, а результаты валидации усредняются по всем циклам. В этом процессе создаются разные варианты обучающих (train) и тестируемых (test) выборок и соотношение доли верных ответов в выборке train приблизительно должно быть равно таковым в выборке test. На каждом цикле валидации за качеством обучения могут следить функции потерь: RMSE, MAE, MAPE и др. [4], одну из которых необходимо предварительно выбирать. Задача Функции потерь: сравнивать доли верных ответов с предсказанными искать градиент Функции потерь, менять в соответствии с ним параметры модели для достижения глобального минимума ошибки.

Если набор входных имеет временную привязку, то перемешивание данных в случайном порядке приведет к потере взаимосвязей наблюдений друг с другом. В этом случае необходимо применять кросс-валидацию на скользящем окне [5] (рис.3), которая обучит модель на небольшом отрезке временного ряда (от начала до некоторого t). Затем делаем прогноз на t+n шагов вперед и считаем ошибку. Далее расширяем обучающую выборку и прогнозируем с t+n до t+2\*n , так продолжаем двигать тестовый отрезок ряда до тех пор, пока не упрёмся в последнее доступное наблюдение. В итоге получим столько частей (фолдов), сколько уместится в промежуток между изначальным обучающим отрезком и всей длиной ряда. Так же модель Хольта-Уинтерса имеет ограничения на значения параметров (α, β, γ) от 0 до 1. Поэтому здесь мы используем алгоритм Truncated Newton conjugate gradient [6], который задаёт ограничения этих параметров на крос-валидацию.

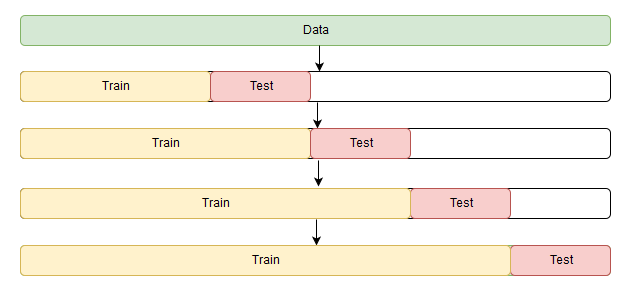


Рисунок 3 – Перекрестная проверка методои скользящего окна.

Исследуемая нами выборка координаты по долготе марки IAT3 содержит около 4 тысяч значений. При этом модель обучилась ~51 секунду и подобрала оптимальные параметры (α, β, γ) и сделала прогноз на 50 шагов вперед (дискретных единиц данных, в нашем случае – дней). . Так же видно, как модель реагирует на изменения ряда затем возвращает дисперсию к обычным значениям Средняя процентная ошибка составляет 6,69% (рис. 4).

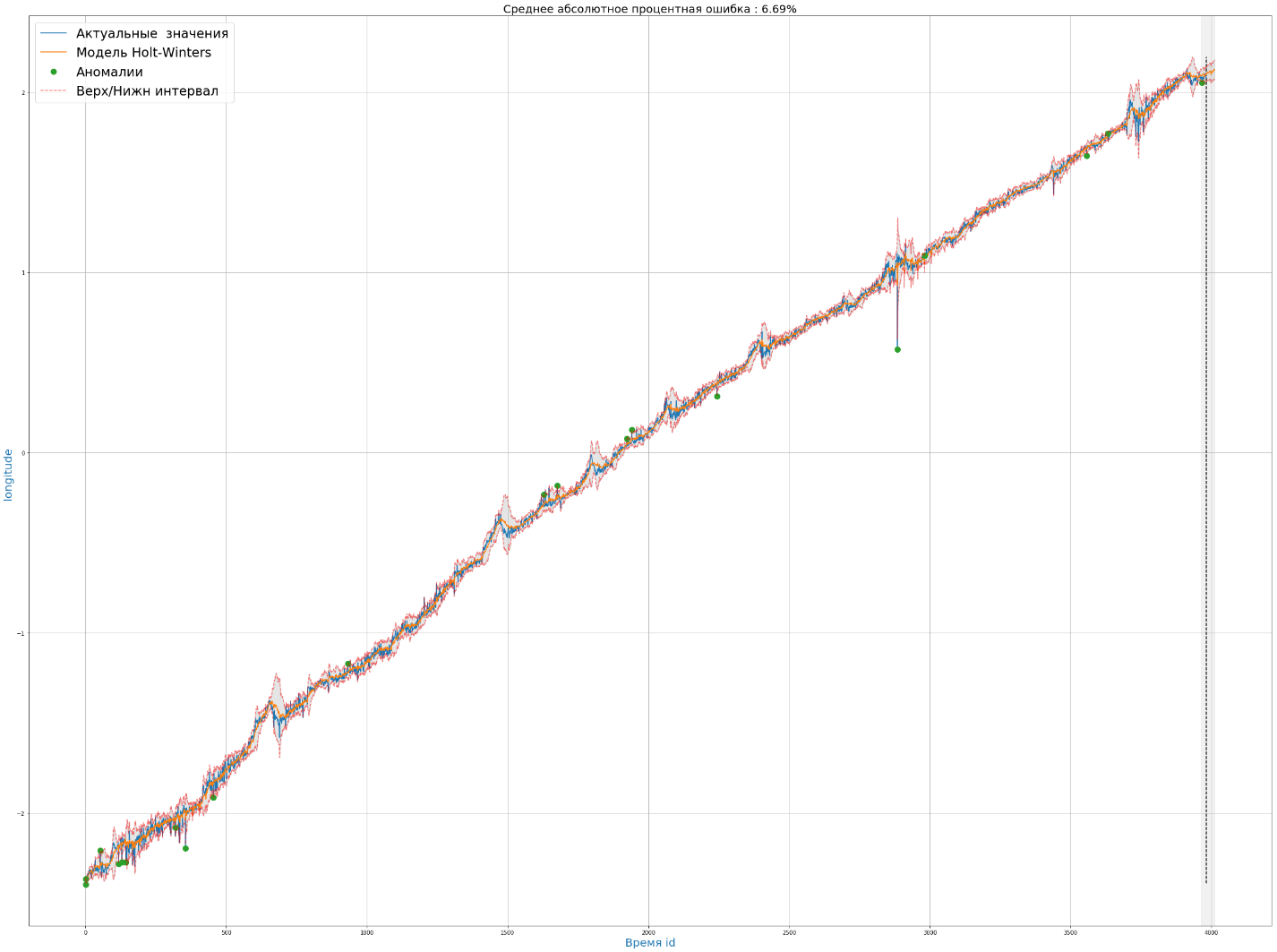


Рисунок 4 – Модель трехпараметрического экспоненциального сглаживания с доверительным интервалом ±3σ., Зеленые точки – идентифицированные выбросы. Горизонтальная шкала - Порядковое количество дней

Следующая модель обучается очень быстро и не требуют слишком больших затрат по подготовке данных – линейная регрессия [7], на вход которой подается одномерный временной ряд координат. При этом необходимо определить признаки, которые подаются на вход обучающей модели. Это могут быть лаги временного ряда, минимальное и максимальное значения анализируемого признака, среднее, дисперсия и т.д. Сдвигая значения координаты отступая, на n шагов назад, мы получаем столбец объектов, в котором текущее значение временного ряда совпадает с его значением во времени t-n, если сделать сдвиг на 1 задержку и обучить модель этой функции, модель сможет прогнозировать на 1 шаг вперед от наблюдения текущего состояния ряда. Увеличение задержки, скажем, до 5, позволит модели делать прогнозы на 5 шагов вперед; однако он будет использовать данные, наблюдаемые 5 шагов назад. Если что-то принципиально изменит ряд в течение этого ненаблюдаемого периода, модель не уловит эти изменения и выдаст прогнозы с большой ошибкой. Следовательно, во время первоначального выбора задержки необходимо найти баланс между оптимальным качеством прогнозирования и длиной горизонта прогнозирования.

На рисунке 5 приведен результат работы линейной регрессии, которая тоже неплохо описывает ряд исходных данных и выделяет выбросы со средней абсолютной ошибкой MAPE 11.03% при скорости обучения 8 мили секунд.

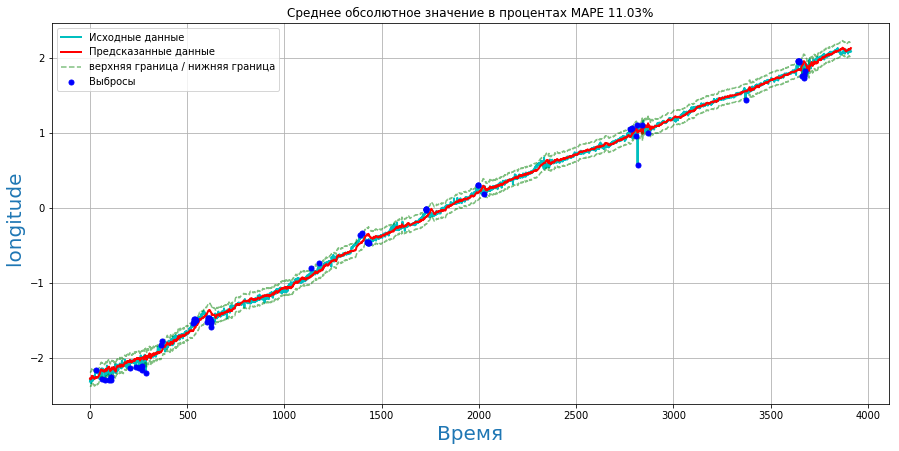
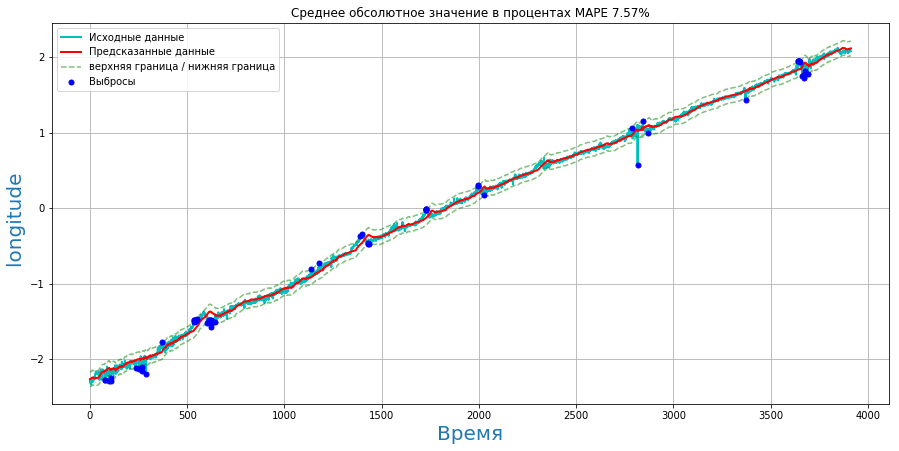


Рисунок 5 – Модель линейной регрессии с доверительным интервалом ±3σ., Синие точки – идентифицированные выбросы. Горизонтальная шкала - Порядковое количество дней

Наиболее популярными регрессионными моделями с регуляризацией являются регрессии Ridge и Lasso [8], которые добавляют еще ограничения к функции потерь и более приближаются к исходным данным. В регрессии Ridge эти ограничения являются суммой квадратов коэффициентов, умноженных на коэффициент регуляризации. Чем больше коэффициент у функции, тем больше будет расхождение модели и входных данных. Следовательно, мы попытаемся оптимизировать модель, сохраняя коэффициенты на достаточно низком уровне.

Рисунок 6 – Ridge регрессия с доверительным интервалом ±3σ., Синие точки – идентифицированные выбросы. Горизонтальная шкала - Порядковое количество дней

Этот Ridge тип регуляризации называется L1, который может уменьшать так же мульти-коллинеарные признаки только что это?. При использовании модели Ridge ошибка упала до 7,57% (рис. 7).

Регрессия Lasso, добавляет к функции потерь не квадраты, а абсолютные значения коэффициентов. В результате в процессе оптимизации коэффициенты несущественных признаков могут становиться нулевыми, что позволяет автоматически выбирать объекты. По модели Lasso нами получены примерно такие же результаты, как и в случае модели Ridge.

Нами опробована еще одна модель XGBoost из , группы алгоритмов градиентного бустинга [9] и деревьев решений. Для корректной работы этой модели нужно удалить тренд, как показано на рисунке 8.

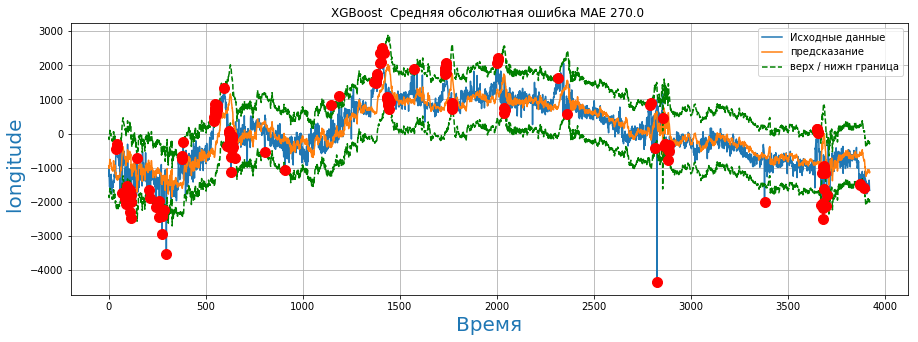


Рисунок 7– работа модели XGBoost. Горизонтальная шкала - Порядковое количество дней

Из рисунка 7 следует, что данная модель достаточно детально описывает поведение входных данных при средней абсолютной ошибке 270 чего. Работа данного алгоритма заняла 30 секунд, что гораздо больше предыдущих моделей, за исключением модели Хольта-Уинтерса.

Так для поиска выбросов и предсказания во временном ряде GPS координат можно применять простые и не затратные модели что бы не тратить время на приведения временного ряда к стационарности для прогнозирования и моделирование эконометрических моделей как SARIMA которая затратила на обучение 1час 24 минуты.

Автор работы выражает благодарность своему научному руководителю к.ф.-м.н., заведующему ЛGPS НС РАН Кузикову С.И. за советы при проведении работ и рекомендации по оформлению статьи.

**Литература**

1. Программная библиотека на языке [Python](https://ru.wikipedia.org/wiki/Python" \o "Python) для обработки и анализа данных [Электронный ресурс] / Pandas documentation

<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.rolling.html>

1. Реализация алгоритма Брутлага в Обнаружении аномалий [Электронный ресурс] / Implementation of Brutlag's algorithm in Anomaly Detection 3. – Режим доступа: https://annals-csis.org/proceedings/2012/pliks/118.pdf (ссылка актуальна на 11.03.2020).
2. Кросс- валидация, [Электронный ресурс] / Time Series cross-validator. – Режим доступа: <https://scikitlearn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html> (ссылка актуальна на 16.03.2020)
3. Выбор метрик качества моделей, [Электронный ресурс] / Regression Metrics.- Режим доступа: <https://medium.com/@george.drakos62/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-1-regrression-metrics-3606e25beae0> (ссылка актуальна на 25.03.2020)
4. Перекрестная проверка на скользящим окне, [Электронный ресурс] / Using k-fold cross-validation for time-series model selection. - Режим доступа: <https://stats.stackexchange.com/questions/14099/using-k-fold-cross-validation-for-time-series-model-selection> (ссылка актуальна на 25.03.2020)
5. Ограниченное число итераций, [Электронный ресурс] / Truncated Newton Method . – Режим доступа: <https://stanford.edu/class/ee364b/lectures/trunc_newton_slides.pdf> (ссылка актуальна на 11.03.2020).
6. Модель линейной регрессии, [Электронный ресурс] / Linear Models . – Режим доступа: <https://scikit-learn.org/stable/modules/linear_model.html> (ссылка актуальна на 16.03.2020).
7. Ridge-Lasso регрессия, [Электронный ресурс] / Regularization: Ridge, Lasso and Elastic Net.- <https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net> (ссылка актуальна на 16.03.2020).
8. Модель XGBoost, [Электронный ресурс] / XGBoost Documentation. – Режим доступа: <https://xgboost.readthedocs.io/en/latest/> (ссылка актуальна на 16.03.2020).

Прохоров Олег Александрович, НС РАН ЛGPS, инженер , [helgpro@yandex.ru](mailto:helgpro@yandex.ru), г. Бишкек.