

Разведочный анализ датасетов по шести загрязнителям, предоставленных Нидерландами, Данией и Сербией за период с 2013 по 2022 годы

Для Нидерландов (NL) EDA выполнила Александра Насонова

Для Дании (DK) и Сербии (RS) EDA выполнила Адель Ким (Мансур Шамсутдинов)

Выводы по каждой работе были сгруппированы в соответствующие подпункты с названиями, соответствующими кодам стран.

1. Jupyter-файлы EDA

NL:

Jupyter-файл EDA с таблицами, графиками и выводами можно найти по следующей ссылке:

https://github.com/AlexandraNasonova/air_pollution_predict/blob/master/airpollpredictor/eda/aqeda.ipynb

Выводы в данном документе по большей части повторяют выводы в jupyter-файле.

DK+RS:

Jupyter-файл EDA по Сербии и Дании можно найти по следующей ссылке:

https://github.com/AlexandraNasonova/air_pollution_predict/blob/master/aqi_eda/aqi_eda_03_analysis_RS_DK.ipynb

2. Анализ пропущенных, невалидных, повторяющихся значений

NL:

2.1. В датасете не всегда указаны значения концентраций (столбец Concentration). Их доля зависит от загрязнителя и составляет от 3 до 7%.

Строки с неуказанными концентрациями необходимо оставить для сохранения последовательности непрерывной.

2.2. Невалидированные данные (Validity ≤ 0 или не определено), для которых концентрация указана, в датасете не найдены.

В соответствии с документацией к датасету, невалидными являются данные за пределами измерений приборов и проч. В данной работы не стоит цель оценки качества работы приборов, поэтому если такие данные будут найдены в датасетах по другим странам, то концентрации необходимо будет заменить на NaN и рассматривать впоследствии как отсутствующие.

2.3. Есть неverified или не прошедшие полный процесс верификации данные (Verification $\neq 1$). Их доля зависит от загрязнителя и составляет от 0 (для NO₂) до 24% (для PM_{2.5}). Такие данные не должны входить в формальные отчеты. Но в нашей учебной работе исключать их необходимости нет. Поэтому мы их будем рассматривать как verified данные.

2.4. Минимальные значения Concentration меньше нуля или равны нулю. Такие концентрации не имеет физического смысла, но, в соответствии с изученными регламентами, можно утверждать, что в датасетах допускается небольшое количество отрицательных значений (если так отработал датчик, а показания должны быть сданы).

Доля отрицательных значений в основном не от 0 до 2,7 %, кроме SO₂ с 9%.

Доля нулевых значений в основном не более 1%, кроме SO₂ с 3%.

Построены распределения отрицательных значений. Пики этих распределений находятся вблизи нуля, но встречаются значения существенно меньше нуля. Последние значения принято решение считать невалидными.

Таким образом, значения вблизи нуля требуется привести к нулю, считая приборной погрешностью. Значения существенно выше нуля, можно считать невалидными и проставить для них NaN.

Что касается SO₂, далее в ходе анализа был построен временной ряд по одной из станций измерения SO₂. И на этом временном ряду видно, что до 2018 года эта станция выдавала очень много отрицательных концентраций. После 2018 года временной ряд стал совершенно другим, без заметных на графике отрицательных концентраций. Видимо поменяли оборудование или еще каким-то образом повлияли на качество измерений. Поэтому, возможно имеет смысл исключить данную станцию из прогноза.

2.5. Были найдены 2 строки с неуказанными значениями для кода станции и даты измерения. Эти строки были удалены, их никак невозможно учесть в прогнозе.

2.6. Для всех загрязнителей, кроме PM_{2.5} и PM₁₀, имеется только один тип временного интервала измерения концентраций - по часам (столбец AveragingTime). В случае с PM_{2.5} и PM₁₀ вторым интервалом является день, других интервалов нет.

С целью сохранения временной последовательности непрерывной и однотипной, было принято решение дополнить датасет строками на каждый час на день до DatetimeEnd строки с интервалом day. В новые строки записать концентрации NaN, а другие значения взять из строки с интервалом day.

2.7. Было обнаружено, что одна станция на один час может выдавать 2 значения концентрации. Эти значения могут быть как одинаковыми, так и различными. Доля таких значений доходит до 30% в случае загрязнителя NO₂. Анализ показал, что более 2х значений не бывает. Дело в том, что одной станции может соответствовать 2 различных значения SamplingProcess. Поскольку эти данные не являются полными дубликатами и несут полезную информацию, то решено их оставить.

DK+RS:

2.8. Первичный анализ данных по совместному датасету Дании и Сербии показал, что в данных по концентрациям есть пропуски, отрицательные значения и очень большие по модулю значения. Также что для всех загрязнителей больше всего данных приходится на станции расположенные в городах (urban в колонке «AirQualityStationArea»). Также оказалось, что в датасете есть по разному агрегированные данные («AveragingTime») - hour, day, var.

2.9. В ходе дальнейшего анализа выяснилось, что в датасете есть невалидные данные, причем обнаружилась прямая корреляция пропусков и невалидных данных. Доля таких данных составляла примерно 6% от общего датасета. Невалидные данные были удалены из датасета, что одновременно избавило нас и от пропусков.

2.10. В нашей работе нас больше всего интересовали данные с «AveragingTime» hour. Данные агрегированные в другом временном интервале встречаются не во всех станциях и не имеют какой-либо ценности для нас, так как они рассчитаны на основе имеющихся данных hour. Дополнительная проверка через DatetimeDelta позволила выявить еще 350 строк с ошибочно указанным AveragingTime равным hour (хотя на самом деле Delta равна 1 дню). Все строки с AveragingTime отличным от hour были исключены из датасета.

В данном случае, возможно, исключение таких данных было не самым оптимальным решением. Возможно стоило их как-то обработать и перевести в часовой формат или наоборот при построении timeseries все данные временно перевести в дневной формат. Так как при построении timeseries во многих станциях были большие промежутки в датах.

2.11. В датасете есть невалифицированные данные. Встречаются они повсеместно и доля таких данных высока, особенно в данных за последние несколько лет, поэтому исключить такие данные мы не можем. Оценка характеристик валифицированных и невалифицированных данных не выявила каких-либо существенных отличий в плане качества данных.

2.12. Оценка долей отрицательных значений показала что на Данию приходится существенно больше таких значений чем на Сербию. После анализа отрицательных значений выяснилось, что основная их часть располагается в довольно узком диапазоне недалеко от нуля, но среди них, также как и с положительными данными, встречаются выбросы, расположенные сильно далеко от основной группы данных. Изначально рассматривалось три разных способа устранения проблемы с отрицательными данными:

- 1) Удалить такие данные, или удалить станции в которых они встречаются
- 2) Заменить их на модуль их значения
- 3) Заменить их на ноль или близкое к нулю значение.

Каждый вышеупомянутый вариант имеет свои недостатки. Так например удалив такие данные мы бы создали пропуски в timeseries, которые пришлось бы как-то восстанавливать. Вариант удалить конкретные станции оказался не очень применим, т.к такие данные встречаются в большинстве станций. Другим возможным решением могло бы быть замещение отрицательных значений их модулями. В данной работе даже было проверено, что для большинства отрицательных значений их модули не превосходили верхние отсекающие границы (усы боксплота) общих выборок каждого загрязнителя. А учитывая их небольшую долю относительно основной выборки, мы бы, наверное, не сильно изменили характеристики основной выборки.

После чтения доступной информации по особенностям работы оборудования измерительных станций выяснилось, что такой способ исправить отрицательные данные был бы скорее всего не совсем корректным. Оказалось, что отрицательные значения могут возникать в связи с плохо откалиброванными детекторами, а также в тех случаях, когда концентрация измеряемого загрязнителя меньше или близка к нижнему пределу обнаружения прибора. Большое количество отрицательных значений в течение продолжительного времени может свидетельствовать о систематической проблеме с качеством калибровки оборудования данной станции. А это уже, в свою очередь, ставит под сомнение качество и валидность в том числе и положительных измерений. Несмотря на это, так как в выбранных странах и так относительно немного данных, то было решено оставить их в датасете и заменить на число близкое к нулю — 0,0001.

2.13. В данных по Дании для некоторых станций были обнаружены дубликаты измерений с одинаковым timestamp'ом. Для Сербии дубликаты не выявились. Однако после чтения информации об особенностях работы измерительных станций, выяснилось, что они могут иметь по два детектора для загрязнителей. И возможно такое удвоение связано с тем, что экспортированные данные с таких станций содержали измерения обоих детекторов по отдельности. Что, впоследствии и подтвердилось после просмотра колонки "SamplingProcess".

После проверки оказалось, что для тех станций, где приведены строки с разными SamplingProcess, но одинаковыми timestamp'ами, во всех случаях значения полностью дублируют друг друга. Поэтому в дальнейшем такие дубликаты были исключены.

3. Анализ выбросов

NL:

3.1. Анализ выбросов был проведен в по каждой станции в отдельности. На boxplot-диаграммах видно, что есть станции на которых выбросы существенны, на 1-2 порядка превышают средние значения.

3.2. Доля выбросов в разрезе загрязнителей составляет от 1 до 7% (SO₂).

3.3. Есть станции с выбросами значительно большими, чем на других станциях. Причем и количество таких выбросов не единично, что может свидетельствовать о том, что это часть их технологического процесса, а не случайность. В таком случае, отбрасывать их нельзя, возможно удастся выделить сезонность.

3.4. Наиболее значительны выбросы у PM_{2.5}, PM₁₀, SO₂ (на 2 порядка). Данные загрязнители являются продуктом работы промышленных предприятий и вполне могут соответствовать реальным значениям, связанным с нештатными ситуациями. Их необходимо учитывать и рассмотреть отдельно.

3.5. Как правило, концентрации в разрезе станций имеют положительно скошенное распределение с одним пиком. Значения среднего и медианы близки, медиана немного меньше.

3.6. У O₃ имеется второй пик вблизи нуля.

DK+RS:

3.7. Была рассчитана доля выбросов каждого загрязнителя в датасете по Сербии и Дании. Для большинства загрязнителей доли выбросов были относительно невысокими (0.35%-8.7%). Однако для SO₂ и CO, рассчитанные доли выбросов оказались равными 23% и 74% соответственно. Сербия была выбрана нами как пример страны с низким качеством воздуха, и вполне возможно, что полученные доли не являются аномалиями а реально связаны с периодическими выбросами высоких концентраций загрязнителей в воздух. Оценка доли загрязнителей в разрезе типов станций, в разрезе их расположения, существенным образом не изменила рассчитанные доли выбросов. Также индивидуальный анализ боксплотов отдельных станций показал, что в некоторых случаях действительно были станции сильно отличающиеся от других по количеству выбросов. Но, к сожалению, на данном этапе не удалось найти оптимального решения по корректировке таких выбросов. Поэтому дальнейший анализ проводился на данных в том виде в каком они были, без удаления и обработки выбросов.

4. Анализ корреляций между загрязнителями

NL:

4.1. Видна очень сильная корреляция между PM_{2.5} и PM₁₀, что понятно, потому что у них один источник - сжигание угля и проч. К тому же возможно при измерении PM₁₀ частично учитываются и PM_{2.5}.

4.2. Есть умеренная корреляция между CO и PM2.5/PM10, , что понятно, потому что у них один источник - сжигание угля и проч.

4.3. Корреляция между SO2 и PM2.5/PM10 очень слабая, хотя у них предположительно также один источник.

4.4. Видна сильная концентрация между CO и NO2, что понятно, потому что у них один источник - сжигание дизельного топлива.

4.5. Имеется довольно сильная отрицательная корреляция между O3 и NO2. Связано это с механизмом формирования ground level ozone (O3) из (в том числе) NOx. Таким образом O3 преобразовываясь из NO2 снижает концентрацию последнего.

DK+RS:

4.6. Выявилась сильная корреляция между PM2.5 и PM10, а также чуть менее выраженная корреляция CO с PM2.5 и PM10. Для других загрязнителей корреляции выражены существенно слабее. Попарные корреляции остаются практически такими же, как при построении графиков по данным всего датасета, так и при построении графиков по одной выбранной станции.

5. Анализ в разрезе метаданных по станциям

NL:

5.1. Из метаданных по станциям были взяты 2 классификатора. Первый - это тип местности, где размещены станции: rural (сельскохозяйственная) и urban (урбанизированная), suburban (промежуточная). Второй - направленность станции: background (нет доминирующих источников), industrial (источники - заводы, теплостанции и проч) и traffic (источники – транспорт).

5.2. Анализ концентраций в разрезе типа местности.

5.2.1. В среднем концентрации O3, PM2.5, PM10 не зависят от типа местности.

5.2.2. Концентрации CO в урбанизированной местности немного выше, чем в сельской. По suburban местности данных нет.

5.2.3. Концентрации SO2 максимальна в урбанизированной местности (в 2 раза выше, чем в сельской), что вероятно связано с наличием там заводов и тепловых станций. Почему в suburban местности средняя концентрация почти в 2 раза меньше, чем в сельской, неясно.

5.2.4. Концентрации NO2 максимальна в урбанизированной местности (в 1.5 раза выше, чем в сельской), и немного выше в suburban. Что вероятно связано с большим количеством транспорта в них.

5.3. Анализ концентраций в разрезе направленности станций.

5.3.1. Концентрации O3 для станций background несколько выше, но существенно для разных направленностей станций не отличаются.

5.3.2. Концентрации PM2.5, PM10, CO существенно для разных направленностей станций не отличаются, в industrial / traffic немного выше.

5.3.3. Данных о SO₂ для станций traffic нет. Для industrial в 3 раза выше, чем для и background.

5.3.4. Концентрации NO₂ в среднем почти в 2 раза выше для станций traffic, чем для background, и почти в 1,5 раза выше для станций industrial, чем для background

DK+RS:

5.4. SO₂: Для Сербии каких-либо существенных отличий как в разрезе типов местности, так и в разрезе типов станций выявлено не было. В категории industrial наблюдается незначительно более высокое медианное значение концентрации. Точно так же rural немного превышает категории suburban и urban. Для Дании для данного загрязнителя есть только 1 тип станций.

5.5. PM₁₀: Для Сербии в категории suburban медианное значение концентрации выше, чем в urban. В разрезе типов станций заметно выделяется категория industrial. traffic и background имеют примерно одинаковые медианные значения. А для Дании наблюдается несколько другая картина: категория urban превосходит rural-regional, а в категории типа станции из имеющихся двух traffic почти в два раза превосходит background.

5.6. O₃: Наблюдается очень сильный переко в сторону rural относительно urban (больше чем в два раза) для датасета Сербии. При этом данные для категории suburban отсутствуют. Возможно это связано с тем, что в Сербии большинство предприятий, вырабатывающих данный загрязнитель располагается в удаленных от городов местах. При этом в разбивке по типам станции отсутствует тип industrial, а медианное значение background станций почти в 5 раз выше чем в traffic. Для Дании наблюдается более ровное распределение, в urban немного ниже чем в rural и rural-regional. А в разрезе типов станций из имеющихся двух категорий медианное значение в background существенно превышает traffic.

5.7. NO₂: Для Сербии наибольшая медианная концентрация наблюдается в urban, она больше чем в два раза выше чем в suburban. А suburban, в свою очередь, больше чем в два раза чем в rural. Т.е. прослеживается довольно четкая зависимость медианной концентрации от удаленности станций от городов. Аналогично и для Дании, наблюдается плавное поэтапное повышение при переходе rural → rural-regional → suburban → urban. Скорее всего это связано с тем, что данный загрязнитель вырабатывается преимущественно при сжигании автомобильного топлива. В разрезе типов станций как для Сербии так и для Дании это косвенно подтверждается тем, что медианное значение traffic заметно превышает другие категории.

5.8. CO: Для обеих стран наблюдается незначительное поэтапное возрастание медианных значений при движении от категории rural к suburban и к urban (хотя для Дании здесь только 2 вида станций rural-regional и urban). А при просмотре типов станций Сербии видно, что в industrial медианное значение примерно на 20-25% ниже чем в других двух категориях(у которых они практически равны). Для Дании получилась более ожидаемая картина, при которой traffic превосходит background.

5.9. PM_{2.5}: Для данного загрязнителя для обеих стран результаты практически полностью перекликаются с таковыми полученными для PM₁₀ (правда различия между разными категориями немного поменьше, но в целом очень похоже), что в принципе и понятно, учитывая их сильную корреляцию и то что они образуются при одних и тех же процессах.

6. Временные ряды концентраций загрязнителей

NL:

6.1. На графике концентрации O₃ видна сезонность с пиками летом и минимума зимой. Вероятно это связано с механизмом образования O₃ через NO_x соединения, которые могут выделяться при аграрных и вегетационных процессах. Преимущественно эти процессы приходятся на лето и высокие температуры. Также на лето приходятся выбросы, причем сравнивая 2 станции складывается впечатление, что выбросы приходятся на одни и те же даты. Возможно это связано с какими-то массовыми сельскохозяйственными работами.

6.2. На графике концентрации PM_{2.5} также видна сезонность, но видна не так явно, требуется проверка. Небольшой подъем приходится на начало года. Поскольку PM_{2.5} активно выделяется при сжигании топлива, то повышенные концентрации могут быть связаны с активным отопительным сезоном, т.е. с периодами низких температур. Так же четко на начало каждого года приходятся огромные выбросы. Есть предположение, что это связано с рождественскими фейерверками.

6.3. График концентрации PM₁₀ очень похож на график PM_{2.5}, что и понятно, потому что эти загрязнители производятся одними и теми же технологическими и природными процессами. Но для PM₁₀ есть большие выбросы на начало года, но есть и другие. Надо посмотреть на какие даты они приходятся. Так же приведен нетипичный график, станция находится на границе города по тем же координатам, что и станция измерения PM_{2.5} с графика с "рождественскими" выбросами, что неожиданно. Может быть оборудование настроено иначе.

6.4. График концентрации CO также на первый взгляд имеет сезонность с подъемами зимой и спадами летом. Вероятно это связано с более активным сжиганием топлива в отопительные сезоны.

6.5. График концентрации SO₂ возможно имеет убывающий тренд. Кажется также, что есть некоторая сезонность с убыванием концентрации летом. На одном из графиков видно, что внезапно изменилась структура данных, стало меньше отрицательных значений. Возможно сменили оборудование.

6.6. На график концентрации NO₂ видно, что выбросы невелики. Это можно объяснить тем, что основным источником NO₂ являются дизельные двигатели и ситуаций, когда их начинают внезапно использовать намного больше обычного, нетипичны. Возможно есть сезонность с подъемом зимой и спуском летом, что можно также объяснить особенностями работы двигателей при низких температурах. Данных после 2018 года нет, объяснения этому пока не найдены.

DK+RS:

6.7. Для усредненному timeseries по всем станциям Сербии хорошо видна сезонность для 1(SO₂), 5(PM₁₀), 7(O₃), 10(CO), 6001(PM_{2.5}). Для 8(NO₂) тоже вполне можно уловить определенную сезонность/цикличность, но чуть менее выраженную, чем для других загрязнителей. Помимо timeseries усредненным по всем станциям, были построены timeseries на основе отдельных выбранных станций с наибольшим количеством данных, по одной станции на каждую страну и загрязнитель.

6.8. Для 1(SO₂) для Сербии не очень хорошо прослеживаются какие-либо закономерности. У Дании чуть получше, вроде бы даже имеются какие-то намеки на сезонность/цикличность, но из имеющейся картинки сложно сделать какие-то выводы.

6.9. То же самое для 5(PM₁₀). У Сербии для выбранной станции представлены неполные данные. Но несмотря на это, можно увидеть сезонное цикличное повышение в декабре-январе. Для Дании вроде бы тоже можно уловить какую-то цикличность, но опять же по приведенным timeseries сложно делать какие-то серьезные достоверные выводы.

6.10. Для 7(O₃) для обеих стран имеется четко прослеживаемая сезонность, с пиками в середине года.

6.11. Для 8(NO₂) для Сербии timeseries не очень хороший. А вот для Дании хорошо видна сезонность с максимумами примерно в декабре-январе

6.12. Для 10(CO) для обеих стран прослеживается сезонность, с пиками в конце года. Хотя у Дании сезонность выражена чуть меньше.

6.13. Для 6001(PM_{2.5}) для обеих стран тоже можно уловить сезонность, с пиками в конце года.