Краткий отчет по встрече (18.06.2025)

---

1. Форматы и структура данных

Лабораторные данные (калибровка)

- Применяются для построения и обучения моделей восстановления концентраций газов.
- Включают сигналы с каждого сенсора по каждому газу (`NO2op1`, `COop1`, `O3op1`, `SO2op1`, `H2Sop1` и др.), а также температуры и влажности:
  - **Температура**: с самого датчика газа (`NO2t`, `COt` и др.), с модуля (`MT`, иногда как `T`), а также иногда с внешних датчиков (например, в G3 — `rhtester`).
  - **Влажность**: с модуля (`MH`), с внешних датчиков или станции.
- В этих данных присутствуют **заданные концентрации газов** (контролируемая подача на стенде), однако возможны эффекты кросс-чувствительности — когда датчик реагирует не только на свой газ, но и на следы других газов или продукты кросс-реакции.
- К каждой точке привязаны маски:
  - `*_bl_auto`, `*_stat_auto` — автоматическая разметка baseline/статичных участков для каждого газа.
  - `bl`, `stat` — ручная или общая маска (пересечение по газам).
- Эти маски позволяют выделить участки, пригодные для оценки baseline и калибровки чувствительности (см. ниже).

Полевые данные (улица)

- Используются для валидации и тестирования моделей, а также для анализа деградации и “дрифта” baseline.
- Структура аналогична лабораторным: сигналы оп-1 и оп-2 для каждого газа, температуры и влажности с разных сенсоров, но **истинных концентраций газов обычно нет**.
- В столбцах концентраций (`NO2`, `CO` и т.д.) указаны значения, рассчитанные по “старой” встроенной калибровке (модель, построенная на предыдущих лабораторных данных).
- Полевые данные отличаются высоким уровнем шума, сезонными и суточными вариациями температуры и влажности, проявлением drift baseline и деградацией сенсора.
- Разметка плато (baseline/stat) отсутствует или требует автоматического выделения — вручную такие большие массивы не размечаются.

---

2. Важные технические моменты

- **Выбор признаков для моделей baseline и чувствительности:**
  - Температура и влажность должны быть максимально близкими к сенсору: в G1/G2 — предпочтительно поля типа `NO2t`, `COt` (температура с датчика), в G3 — `rhtester`, для влажности аналогично.
  - Использование менее “близких” параметров (например, только температуры/влажности с модуля) ухудшает точность baseline.
- **Кросс-чувствительность датчиков:**

    - При калибровке на один газ может наблюдаться реакция на другие (следы или продукты кросс-реакции).
    - Для большинства газов (NO2, O3) этим можно пренебречь, но для H2S, SO2 и др. — желательно проверить регрессионной моделью (сигнал датчика vs. все концентрации).
- **Старая калибровочная модель концентраций на полевых данных:**
    - Обычно это простая линейная формула:

      ```
      Conc = (Signal - Baseline(T, RH)) / Sensitivity(T, RH)
      ```

      где параметры Baseline и Sensitivity подбирались по лабораторным данным.
    - Эта модель часто не учитывает деградацию или сезонный drift baseline, поэтому на реальных данных возникают ошибки.
- **Методы оптимизации:**
    - Для нахождения параметров baseline/чувствительности использовались методы наименьших квадратов (МНК), которые хорошо подходят для “чистых” данных, но могут застревать в локальных минимумах, если baseline нестабилен или много выбросов.
    - Пробовали и более продвинутые методы, например, дифференциальную эволюцию (эволюционный стохастический алгоритм, устойчивый к сложным ошибочным поверхностям).

---

3. Drift и деградация сенсоров

- Все модели, обученные только на лабораторных данных, не способны учесть эффекты деградации (старения) сенсора и дрейфа baseline, возникающего при длительной эксплуатации на улице.
- На полевых данных часто наблюдается:
    - Смещение baseline (особенно летом, при высоких температурах и влажности).
    - Завалы baseline вниз, появление отрицательных значений восстановленной концентрации.
    - Рост случайного и структурного шума.
- Для учёта деградации:
    - Необходимо дообучать baseline-модель на уличных данных по выделенным “чистым” плато (участки стабильного сигнала, где газа скорее всего нет).
    - Без внешнего референса — возможно корректировать только baseline, не чувствительность.
    - Если есть референс или эталон — можно строить semi-supervised подход: дообучать и baseline, и чувствительность на участке с известной концентрацией.

---

4. Практические советы и идеи для аналитики

- **Автоматическое выделение baseline на уличных данных:**
    - Использовать скользящую стандартную ошибку (rolling std) для поиска плато: если std сигнала низкая в течение заданного окна — вероятно, это участок baseline.

    - Учитывать совпадения плато по температуре и влажности (для исключения влияния их скачков).
    - Применять мульти-датчиковый анализ: если на большинстве датчиков плато, а на одном есть отклонения — этот датчик, вероятно, в этот момент реагирует на свой газ, остальные — находятся в baseline.
- **Кросс-газовый анализ:**
    - Строить регрессию сигнала каждого датчика по всем поданным концентрациям газов — если коэффициенты при “чужих” газах близки к нулю, перекрёстная чувствительность отсутствует.
- **Semi-supervised дообучение:**
    - Если на полевых данных есть хоть иногда эталонные калибровочные замеры (референс) — использовать их для частичного дообучения модели baseline/чувствительности.
    - В остальное время корректировать baseline только по “нулевым” плато сигнала (например, ночью или при длительных низких значениях).

---

5. Развёрнутый To-Do

1. Проанализировать используемые температуры/влажности в разных файлах и скорректировать выбор признаков для baseline/чувствительности.
2. Разработать или внедрить автоматизированную функцию выделения плато (baseline) на уличных данных — rolling std + анализ по всем датчикам.
3. Оценить и количественно проверить кросс-чувствительность датчиков по лабораторным данным, обучая многомерную регрессию для каждого сигнала.
4. Реализовать алгоритмы корректировки baseline (и при возможности — чувствительности) на уличных данных для борьбы с drift и деградацией.
5. Если есть эталон или внешние калибровочные периоды — интегрировать semi-supervised корректировку baseline и чувствительности.
6. Построить и протестировать стратегию валидации: анализировать распределения сигнала, сравнивать результаты различных моделей baseline/чувствительности, фиксировать “аномалии”.
7. Документировать все ключевые этапы и обозначения, чтобы была единая система соответствий “колонка → газ/датчик/физическая величина”.

> **Если нужны примеры кода для автодетекции baseline, анализа кроссчувствительности, построения semi-supervised pipeline или любые дополнительные пояснения по аналитике — пиши, помогу!**