Messy Energy Data. Sense-making via change-point and anomaly detection

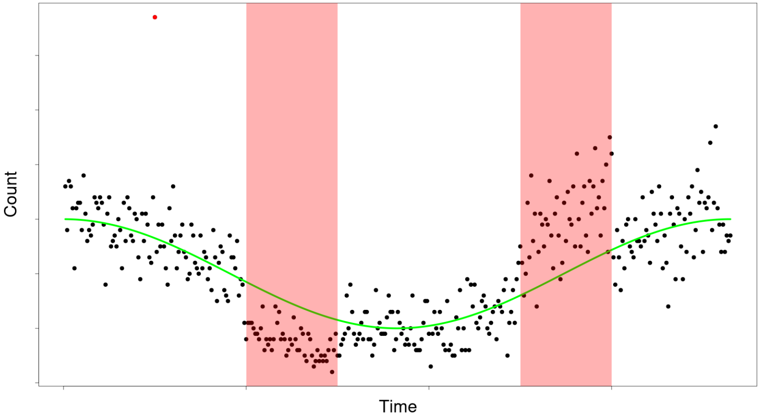
Sept 2024

Paul Smith & Idris Eckley

Acknowledge Colleagues

- Avery Chong
- Christian Remy
- Adam Tyler
- Oliver Bates
- Adrian Friday

Collective and Point Anomalies



Determining Anomalies

Consider the observed data $\mathbf{y}_{1:T} = (y_1, \ldots, y_T)$ with K collective anomalies $\mathbf{y}_{s_1:e_1}, \ldots, \mathbf{y}_{s_k:e_k}$

- The background cost of an observation is $\mathcal{C}\left(y_t\right)$
- An anomalous period $\mathbf{y}_{s:e}$ has parameter perturbation $\hat{\psi}_{s:e} = \min_{\psi} \sum_{t=s}^{e} \mathcal{C}\left(y_t, \psi\right)$ giving cost $\sum_{t=s}^{e} \mathcal{C}\left(y_t, \hat{\psi}_{s:e}\right)$.
- Penalties for introducing point (β_P) and collective (β_C) anomalies that do not depend on K

Identification of Anomalies

Select $K, s_1, e_1, \ldots, s_K, e_K$ by minimising

$$F_T = \sum_{t \notin \cup_{i=1}^{K} s_i : e_i} \min \left\{ \mathcal{C}\left(y_t\right), \mathcal{C}\left(y_t, \hat{\psi}_t\right) + \beta_P \right\} + \sum_{i=1}^{K} \left\{ \sum_{t=s_i}^{e_i} \mathcal{C}\left(y_t, \hat{\psi}_{s_i:e_i}\right) + \beta_C \right\}$$

General dynamic programming solution is $\mathcal{O}\left(n^2\right)$

Identification of Anomalies

Under conditions

- $\min(F_T) \geq \min(F_{T-1})$
- $\exists \kappa$ s.t. $\min(F_T) \leq \min(F_{T-1}) + \kappa$

the solution is $\mathcal{O}(n)$[1]

Satisfied if cost is taken to be the Deviance.

[1] Fisch et al. 2022. https://doi.org/10.1002/sam.11586

LU Campus Energy Data

- 75 Buildings / Building Groups
- 1594 sensors

Substance	Number
Electricity	1028
Gas	71
Water	181
Heat	313
Oil	1

Contextual Data for Meters

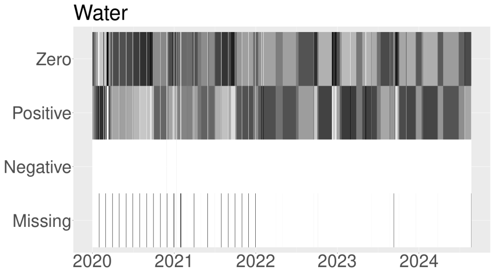
- Location in building
- Textual description of monitored area
  - No record of what was going on in that area
- Unknown hierarchy
- Loggers record as a count every ten minutes
  - variable resolution
- Historically(?) fragile data pipeline

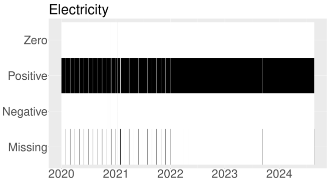
Data Screening

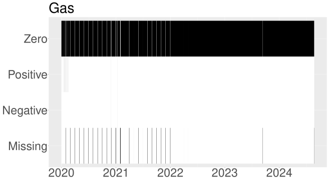
- Assign each observation to one of four Classes
  - Positive
  - Zero
  - Negative
  - Missing

  then aggregate the data to Daily
- Costs derived from the Multinomial distribution
- Background cost is based on parameters representing performance

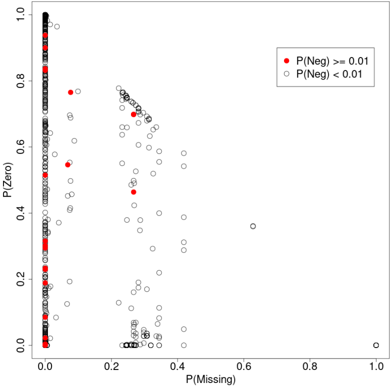
Data Screening (Heatmaps)







Data Screening (Intervention)

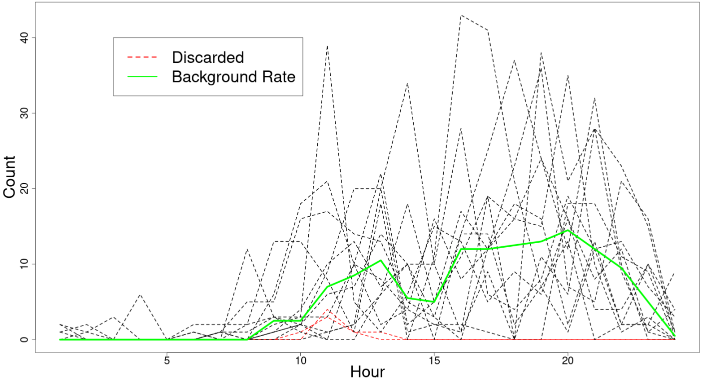


Condition	Num. Sensors
P(Missing) > 0.1	221
P(Zero) > 0.9	514

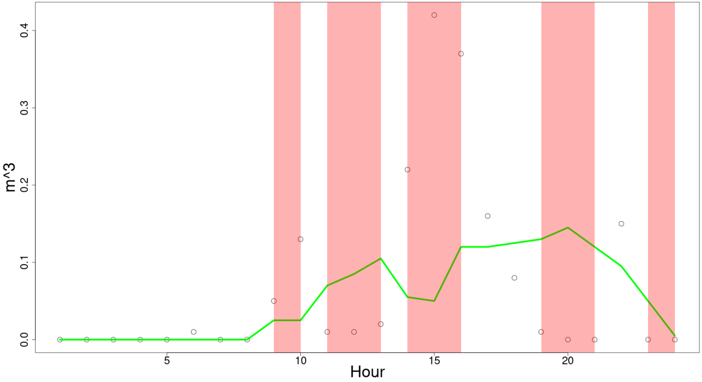
Changes in Daily Usage Patterns

- Uniqueness of place and process
- For day n, use days $n-1, \ldots, n-14$ to build the background distribution
    - Discard anything identified and explained as anomalous
- Treat the data as counts
    - Costs based on the Poisson distribution
- Propose the kind of anomalous change
    - Proportional increase in rate

Estimating the Background

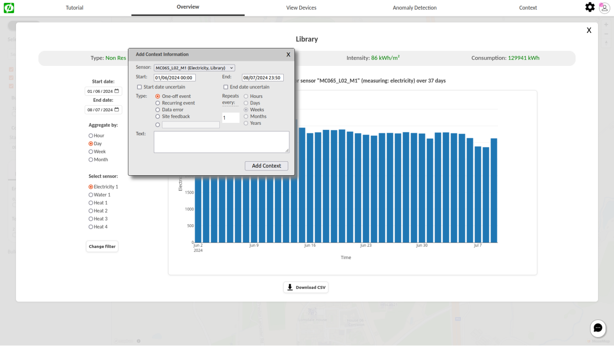


Anomalies



Gathering Context

To learn we need to understand why. . .



Summary

- Introduced an efficient method for detecting anomalies
  - Extensions e.g. multivariate series not covered here
- Outlined some challenges of working with energy data in the wild
- Shown how the anomaly techniques can inform an exploratory data analysis

Thanks!

p.j.smith@lancaster.ac.uk