UNIVERSITATEA BABEȘ-BOLYAI

ANOMALY DETECTION USING TIME SERIES ANALYSIS IN THE VARIATION OF RADON CONCENTRATION

Ş. GRECU, T. DICU, A. CUCOȘ, M. BOTOŞ, G. DOBREI

"CONSTANTIN COSMA" RADON LABORATORY,
FACULTY OF ENVIRONMENTAL SCIENCE AND ENGINEERING,
BABEȘ BOLYAI UNIVERSITY, CLUJ-NAPOCA,
ROMANIA

CURRENT STATE OF KNOWLEDGE

Spatial variation of radon

Geology and soil composition, soil permeability, Uranium content, building construction and ventilation, human behaviour (ventilation patterns) etc.

Temporal variation of radon

High concentration in the cold season and low concentration in the warm season.

Exceptions that do not follow the classic pattern

Studies that seek to determine and explain anomalies in radon concentration variation.

Anomaly detection studies

Statistical methods, machine learning methods (unsupervised, supervised).

References:

- Cosma C., Dicu T., Dinu A., Begy R., 2009, *Radioactivitatea Mediului, Radonul și cancerul pulmonar*, Editura Quantum, Cluj-Napoca.
- Haider, T., Adnan B., Umar, H. ,.. 2020. Identification of radon anomalies induced by earthquake activity using intelligent systems. Journal of Geochemical Exploration. 222. 10.1016/j.gexplo.2020.106709.
- Fujiyoshi, R., Morimoto, H. & Sawamura, S. (2002). Investigation of the soil radon variation during the winter months in Sapporo, Japan. *Chemosphere*. 47 (4), 369–373. https://doi.org/10.1016/S0045-6535(01)00310-1

OBJECTIVES

Standard deviation from the mean

Density-based clustering algorithm (DBSCAN)

Autoencoders

Anomalies

8 houses in Romania

Measurement period: 01.01.2019 - 13.03.2023

Measurement frequency: 2 min

TOTAL 8,240,951 measurements

Active measurements

METHODS AND INSTRUMENTS

- **The monitoring system**, called **SmartRadon - ICA**, developed in the Radon Testing Laboratory "Constantin Cosma"- LiRaCC within the project "**SMART_RAD_EN**".

- **100 devices** placed in 5 big cities in Romania (Cluj-Napoca, București, Timișoara, Iași, Sibiu)

- **Results**: indoor air pollutants (Rn, CO, VOC) and **physical parameters** (CO_2, humidity, temperature); implementation of remedial solutions for 10 houses selected in the project

RESULTS

Table 1. Descriptive statistics of radon concentration measured in 8 houses

Code	N	AM	SD	Min	Max	Median	GM	GSD
1	1024577	476	281	8	1500	457	369	2.29
2	997389	475	396	8	3088	391	279	3.34
3	1011689	725	517	8	4336	634	521	2.51
4	1053124	620	431	8	2220	599	413	2.95
5	1031191	673	459	8	5704	650	468	2.77
6	1057198	205	148	8	944	169	148	2.43
7	1034340	293	205	9	1546	243	219	2.28
8	1031443	286	240	8	1749	227	185	2.83

(Code- the specific code for each house; N = number of measurements; STD = standard deviation; Min- minimum value, Max- maximum value; GM- geomtric mean; GSTD- geometric standard deviation

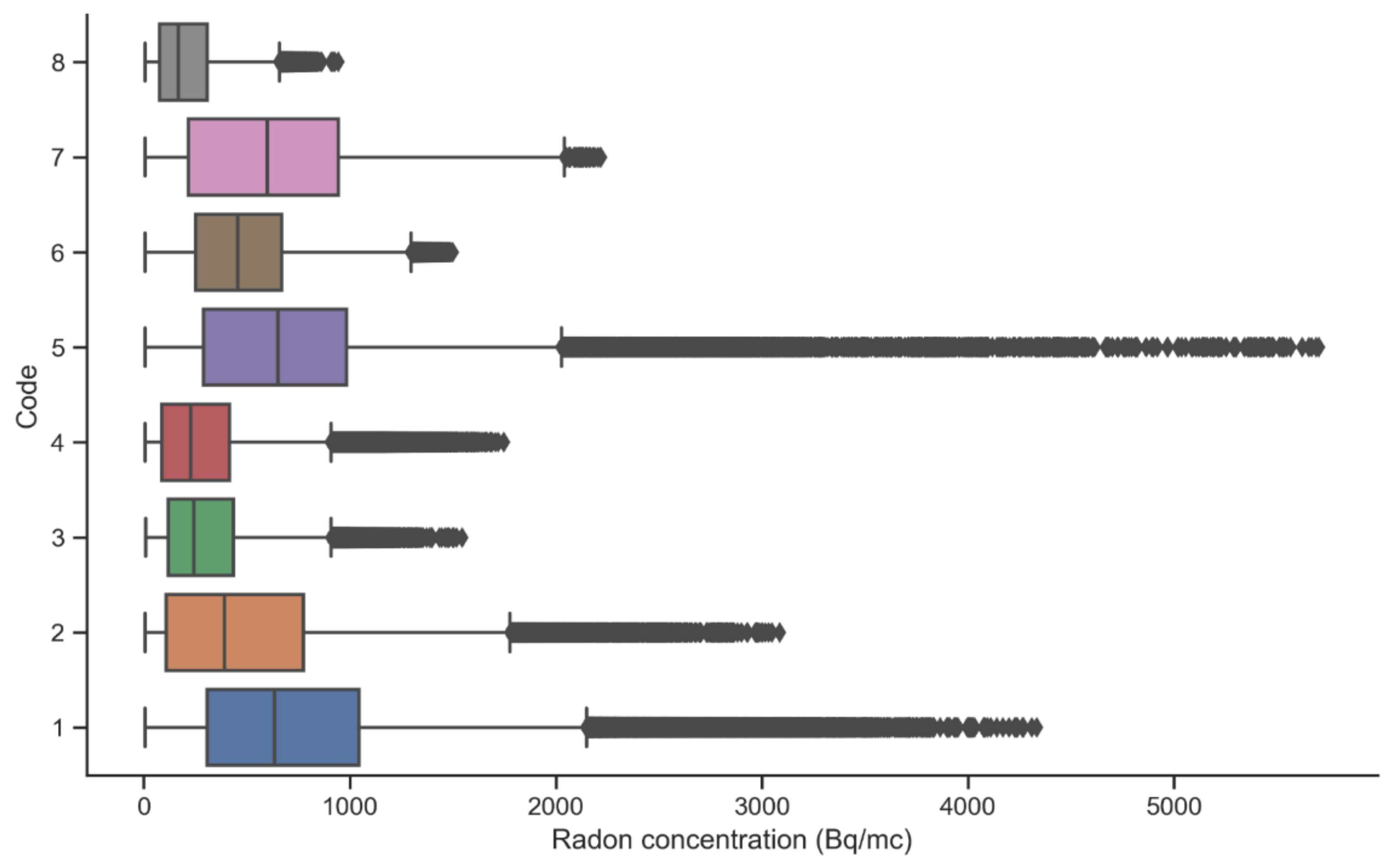


***Figure 1.** Boxplot representation of radon concentration, for the 8 houses*

RESULTS

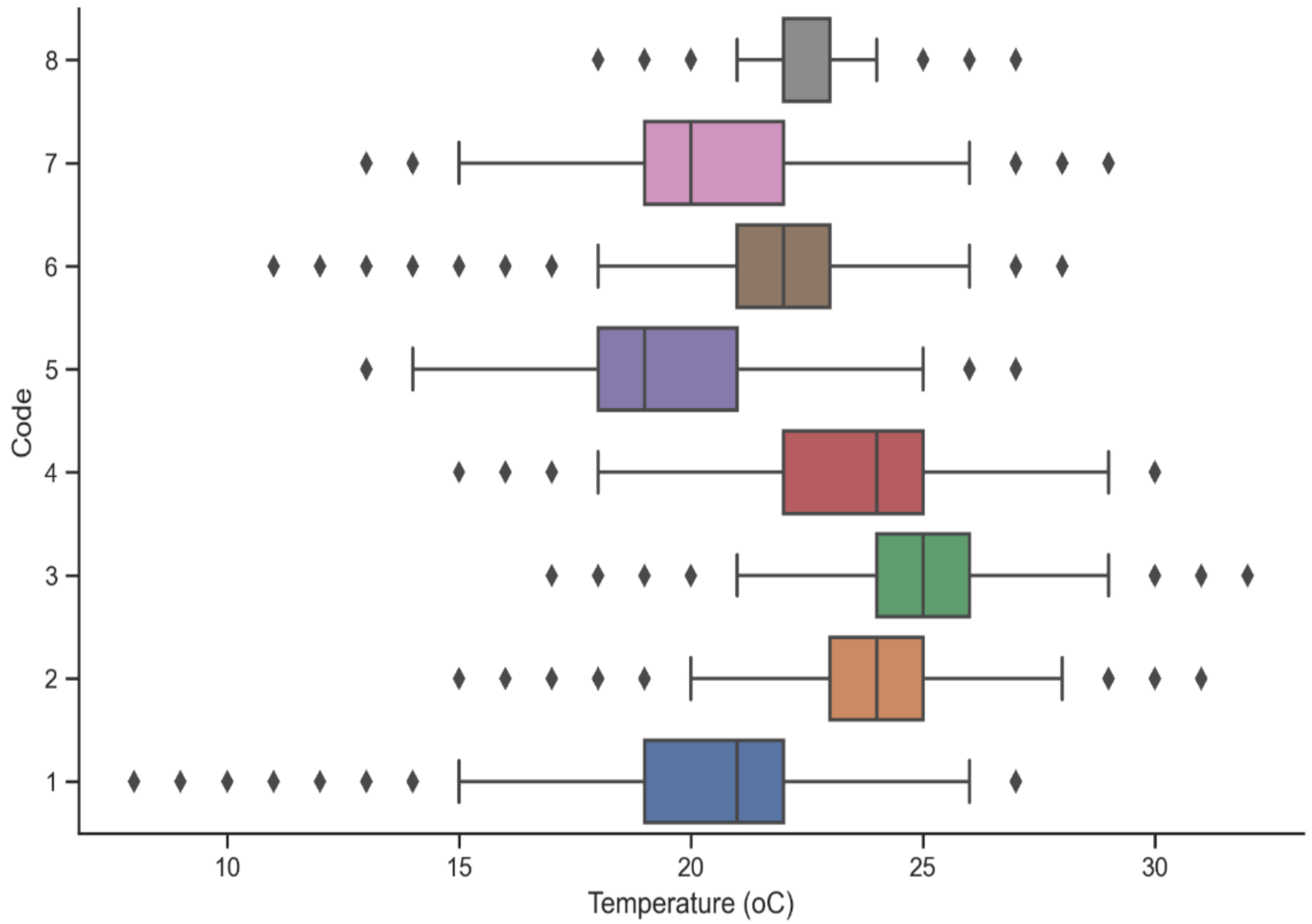


Figure 2. *Boxplot representation of temperature for the 8 houses*

Table 2. *Descriptive statistics of temperature measured in 8 houses*

Code	N	AM	SD	Min	Max	Median	GM	GSD
1	1011689	20.6	1.89	8	27	21	21	1.10
2	997389	24.1	2.22	15	31	24	24	1.09
3	1034340	25.2	1.68	17	32	25	25	1.07
4	1031443	23.7	2.41	15	30	24	24	1.11
5	1031191	19.8	2.45	13	27	19	20	1.13
6	1024577	22.0	1.56	11	28	22	22	1.07
7	1053124	20.4	2.66	13	29	20	20	1.14
8	1057198	22.4	0.88	18	27	22	22	1.04

(Code- the specific code for each house; N = number of measurements; STD = standard deviation; Min- minimum value, Max- maximum value; GM- geomtric mean; GSTD- geometric standard deviation

RESULTS

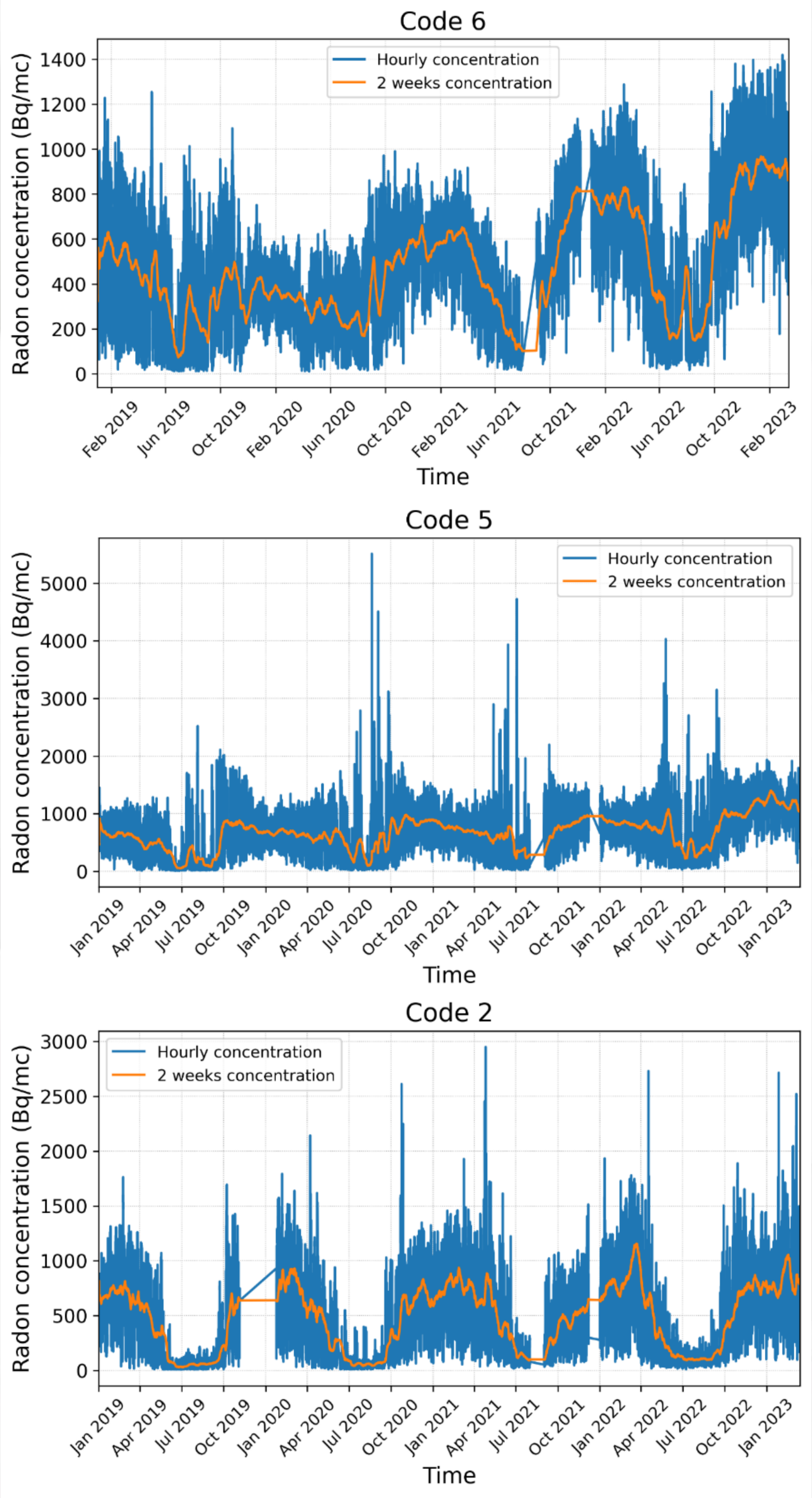


Figure 3. Radon concentration evolution for house code 6 (top), house code 5 (middle) and house code 2 (bottom)

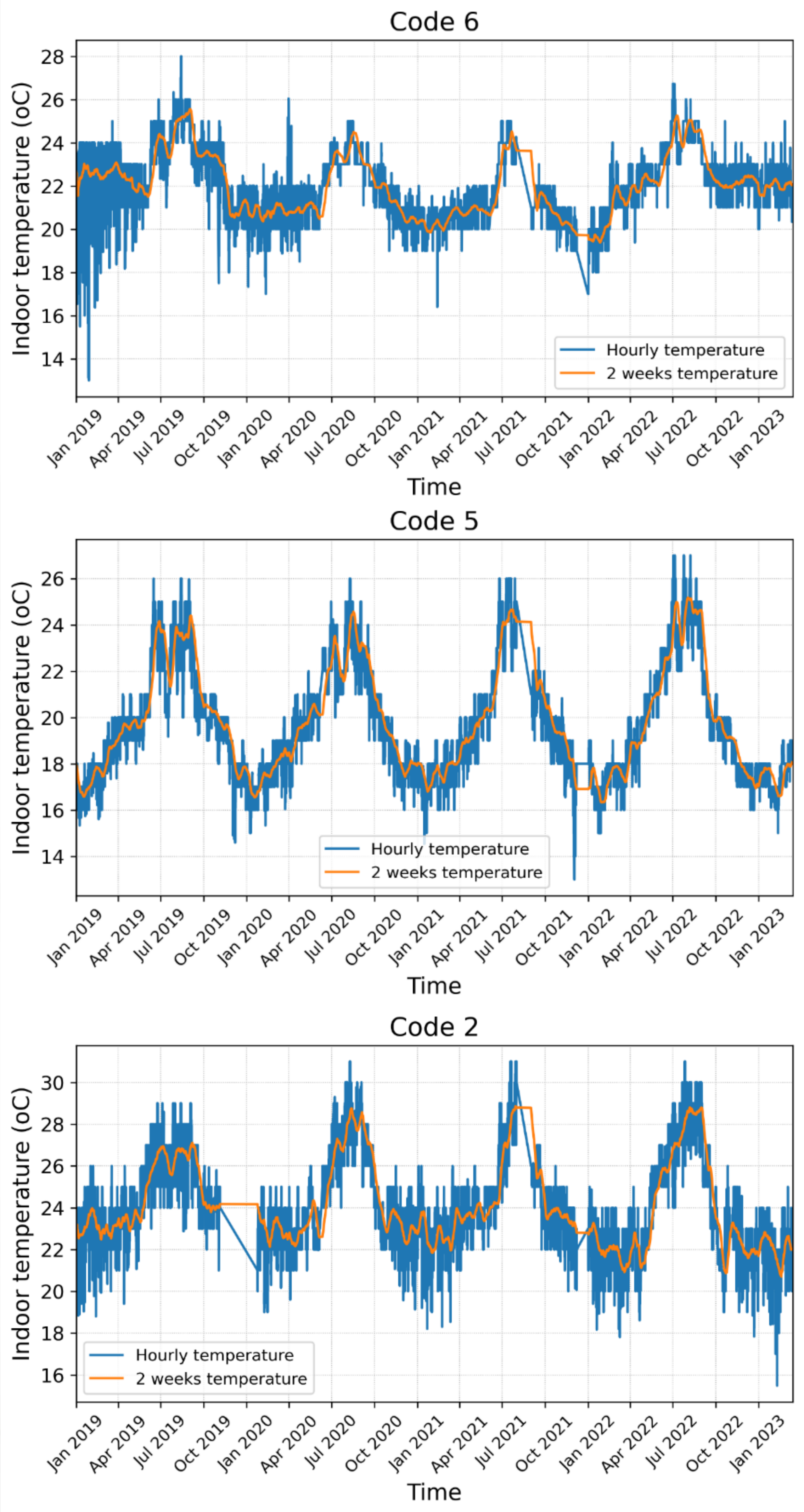


Figure 4. Temperature evolution for house code 6 (top), house code 5 (middle) and house code 2 (bottom)

STATISTICAL METHODS

- Based on data properties for anomaly determination

- Analysis of input data based on their properties → identification of observations/events that deviate significantly from the normal behaviour of the data set

Examples:

- **standard deviation from the mean**- if an observation is within n standard deviations from the mean, it is considered an anomaly; used as individual technique for anomaly detection (Gregoric et al., 2012; Zmazek et al., 2005), or combined with other techniques, as threshold value (Haider et al., 2021; Rafique et al., 2020; Singh et al., 2017).

- Inter quartile range (IQR)

- Regression techniques

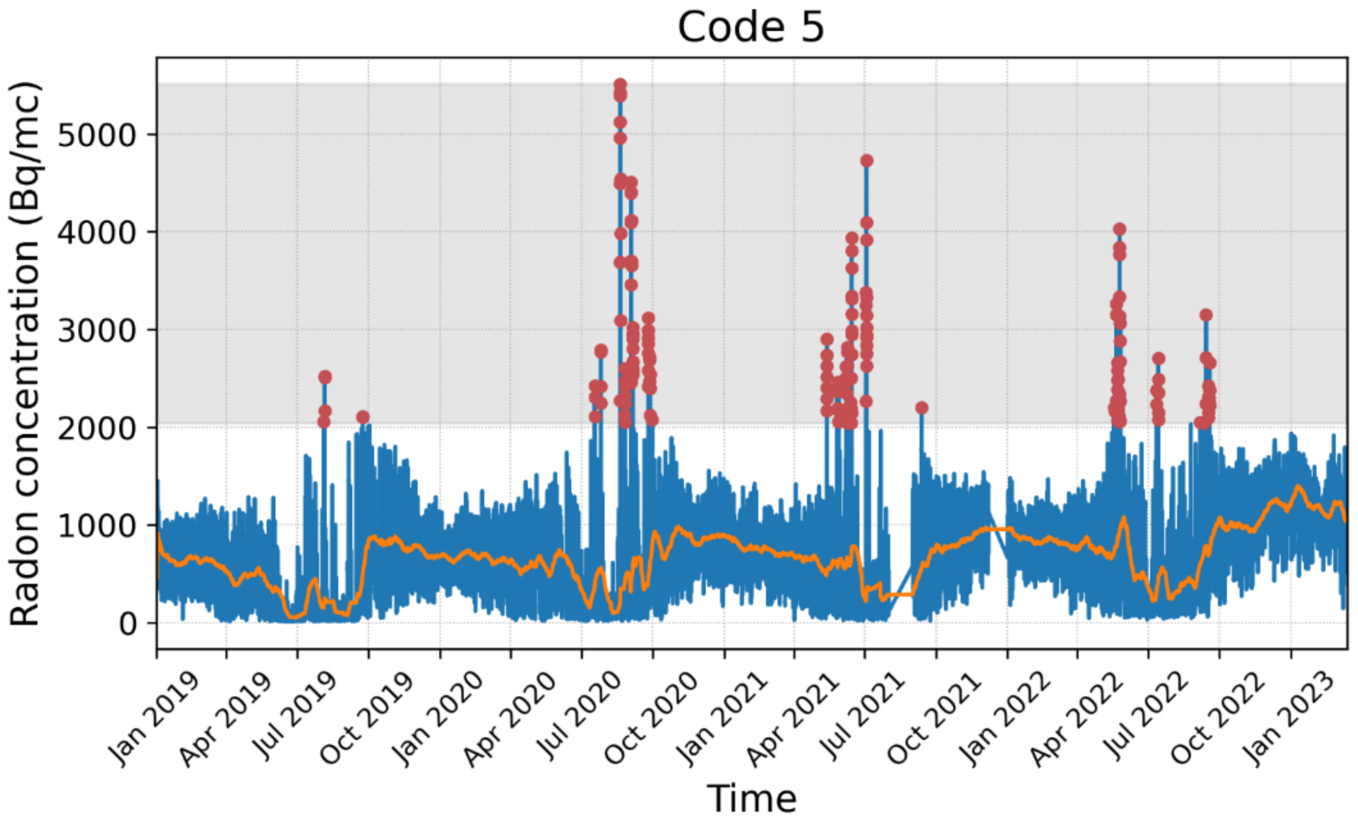
....

Disadvantage of these methods: not taking into consideration the local context of the data; for example, a sudden increase in radon concentration is not classified as an anomaly.

Reference:
Chandola, V., Banerjee, A. & Kumar, V. 2009. Anomaly detection: A survey. ACM Comput. Surv. 41, 3, Article 15 (July 2009), 58 pages. https://doi.org/10.1145/1541880.1541882

RESULTS

- The statistical criterion was applied with parameters (μ+3σ) calculated for the whole study period (4 years and two months), at season level, at month level.



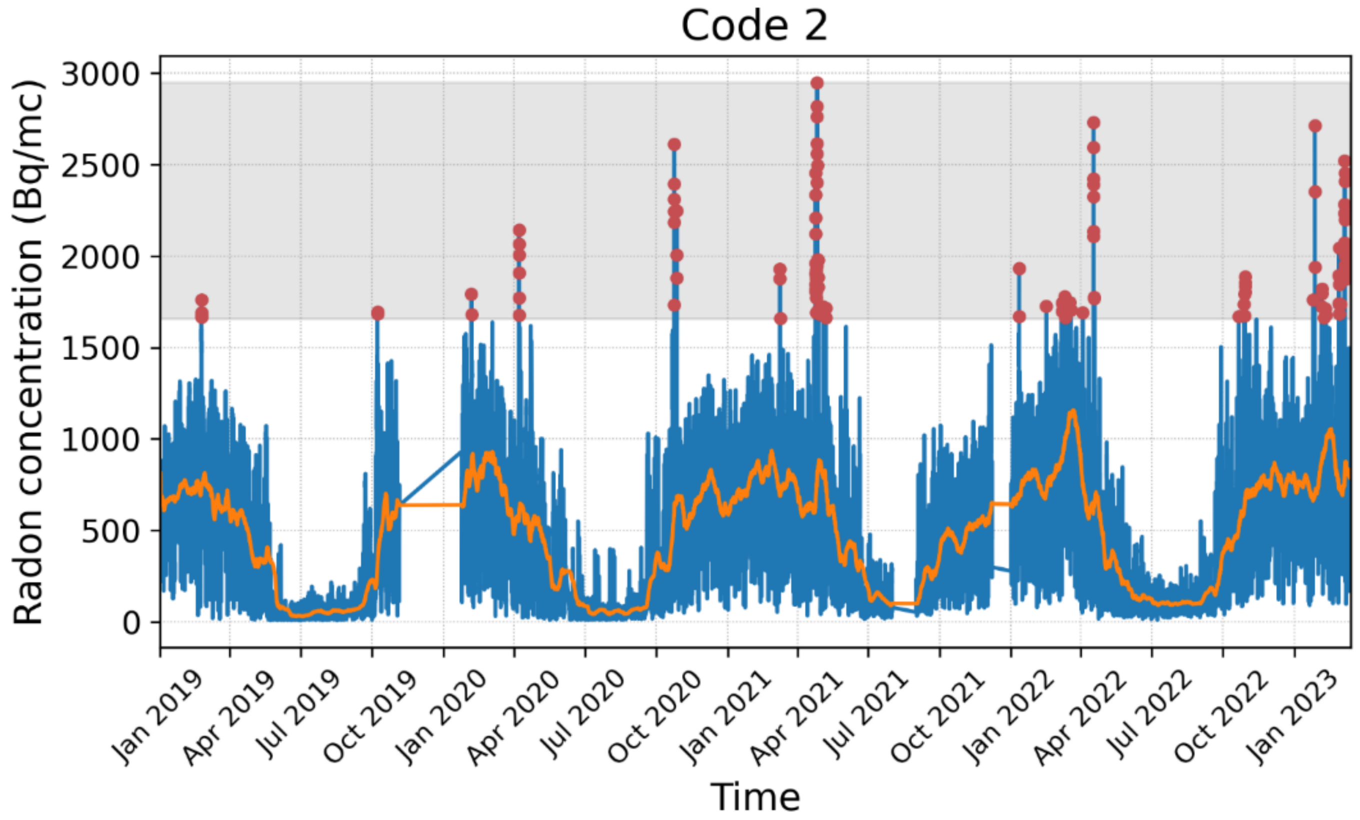
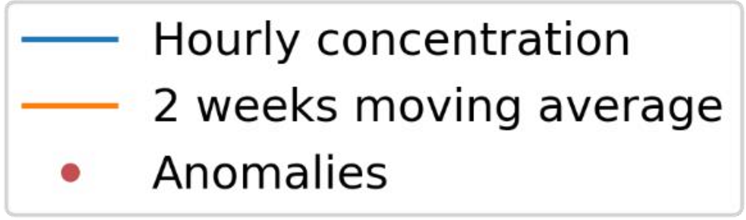
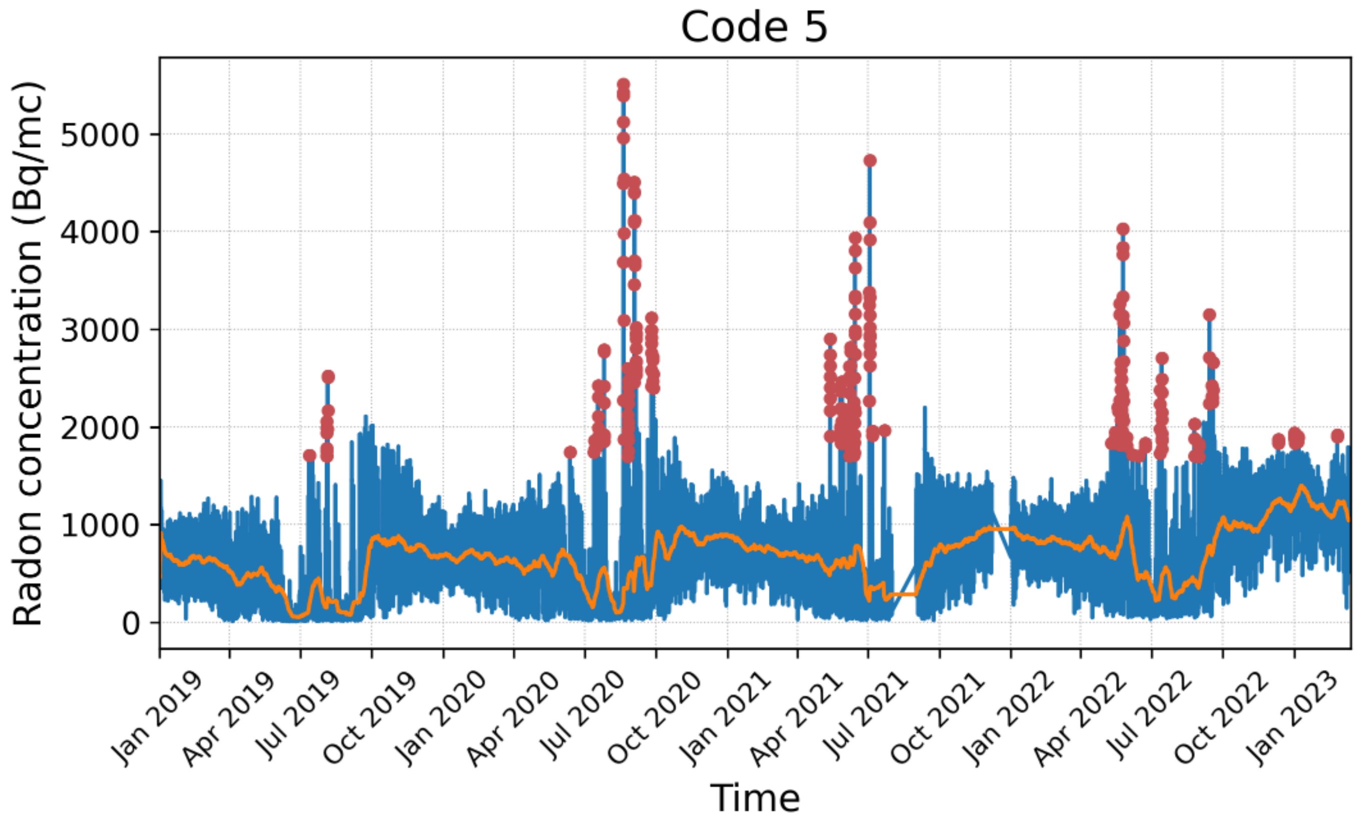


Figure 5. Statistical method (μ+3σ), **with μ și σ calculated for the whole study** period 2019-2023, for houses code 5 (left) and house code 2 (right)





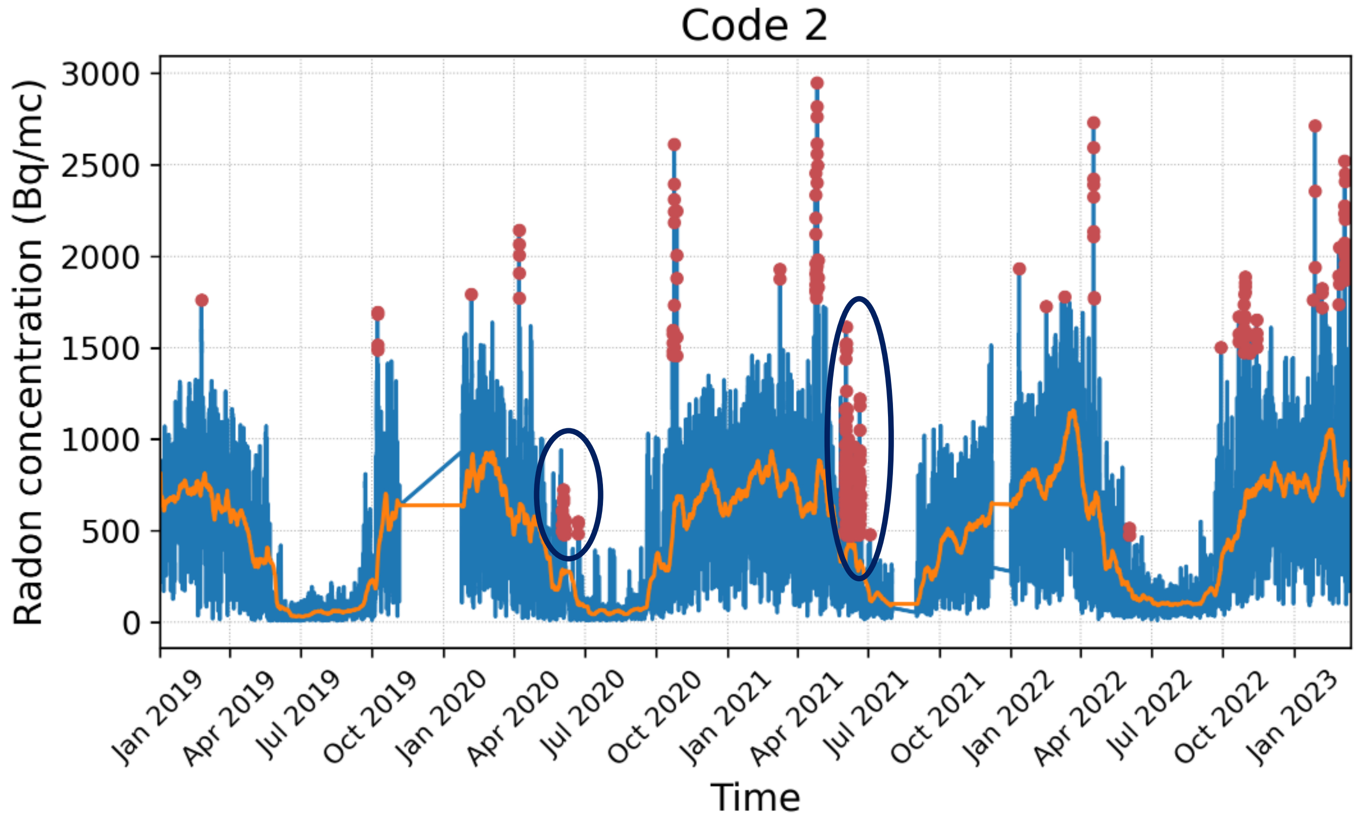


Figure 6. Statistical method (μ+3σ), **with μ și σ calculated at seasonal level**, for houses code 5 (left) and house code 2 (right)

RESULTS

- **Monthy threshold averaging the months across all years**

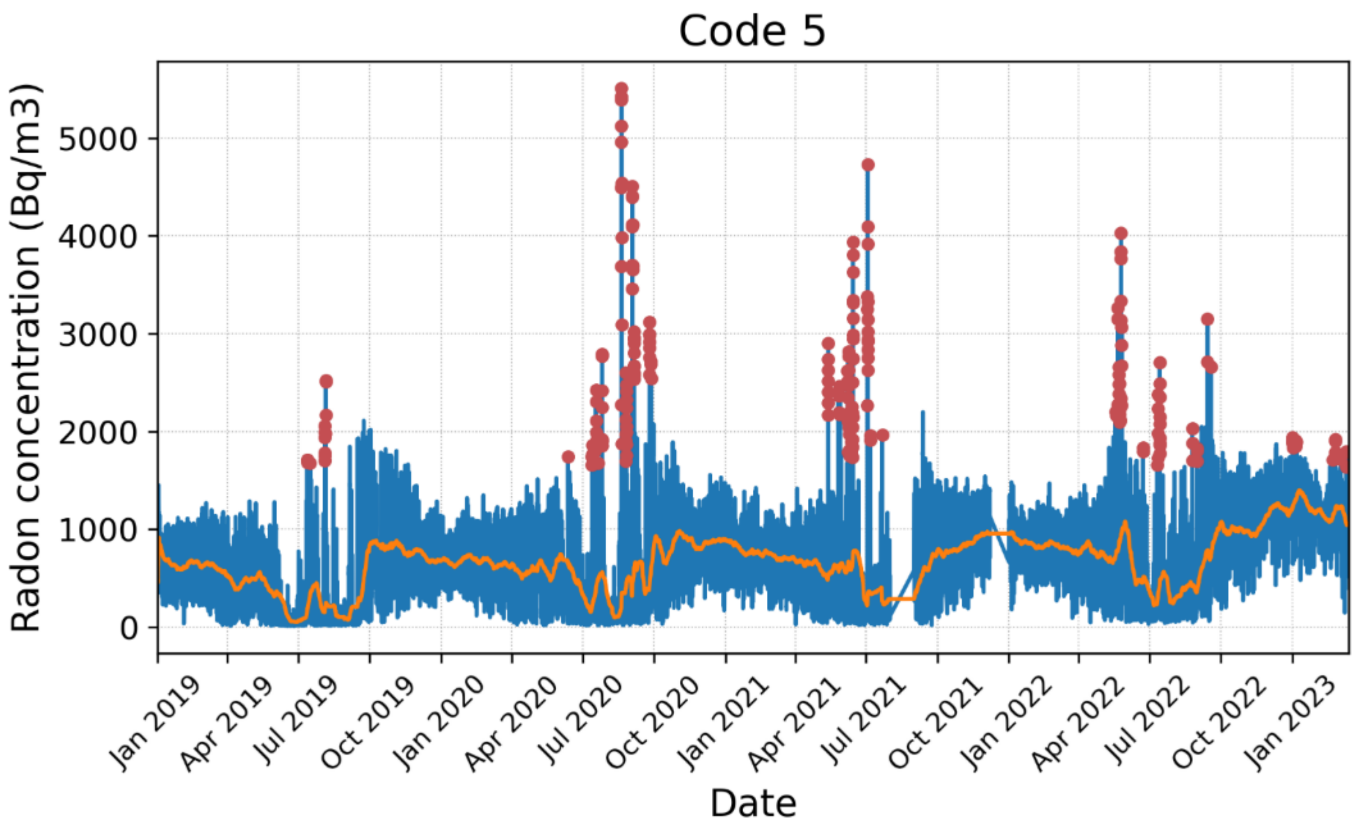
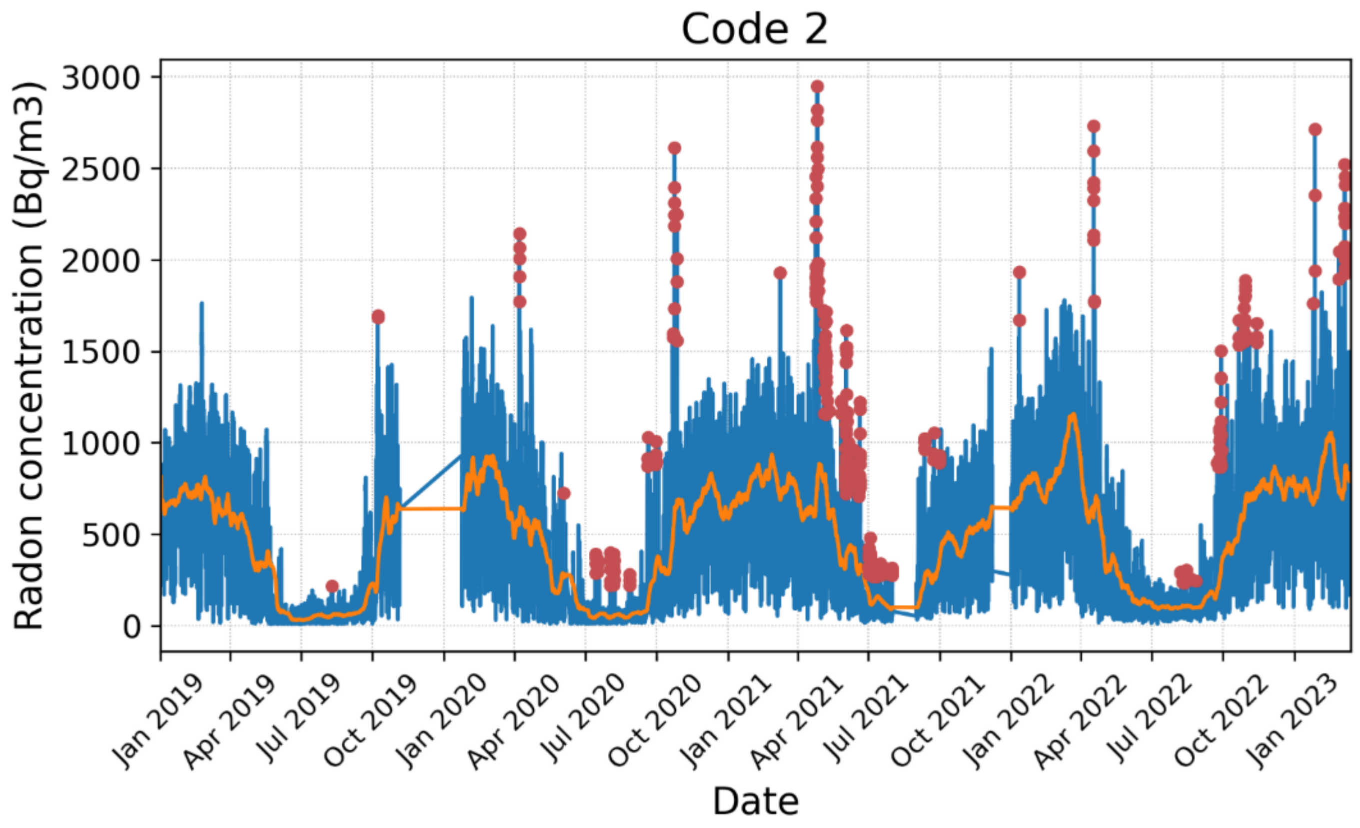
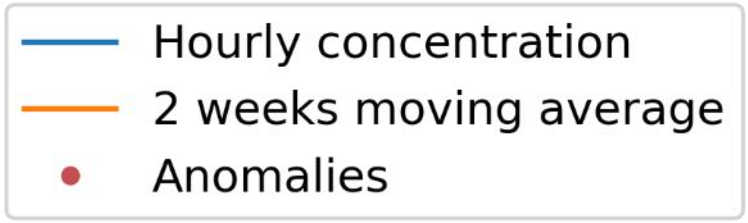


Figure 7. Statistical method (μ+3σ), **with μ și σ calculated averaging the months across all years**, for houses code 5 (left) and house code 2 (right)





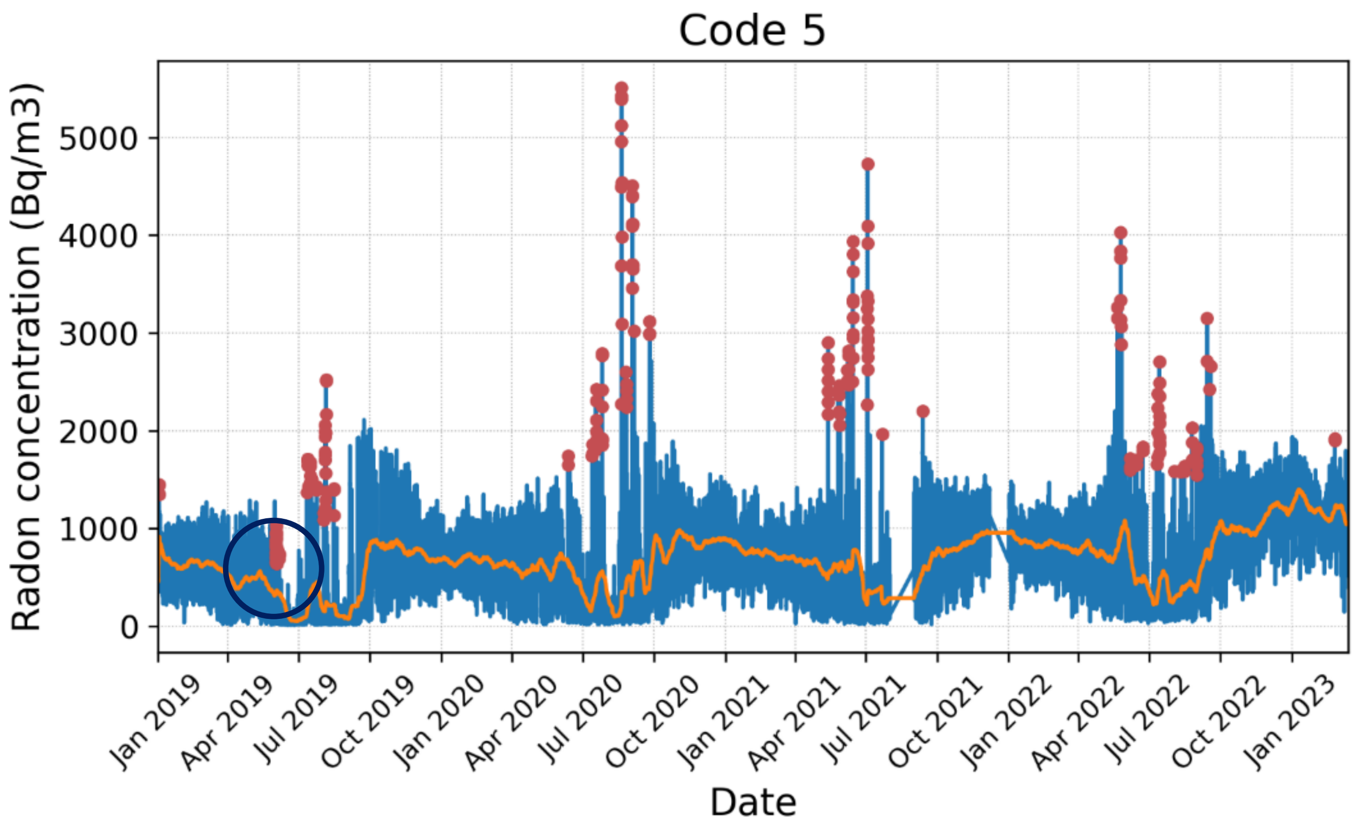
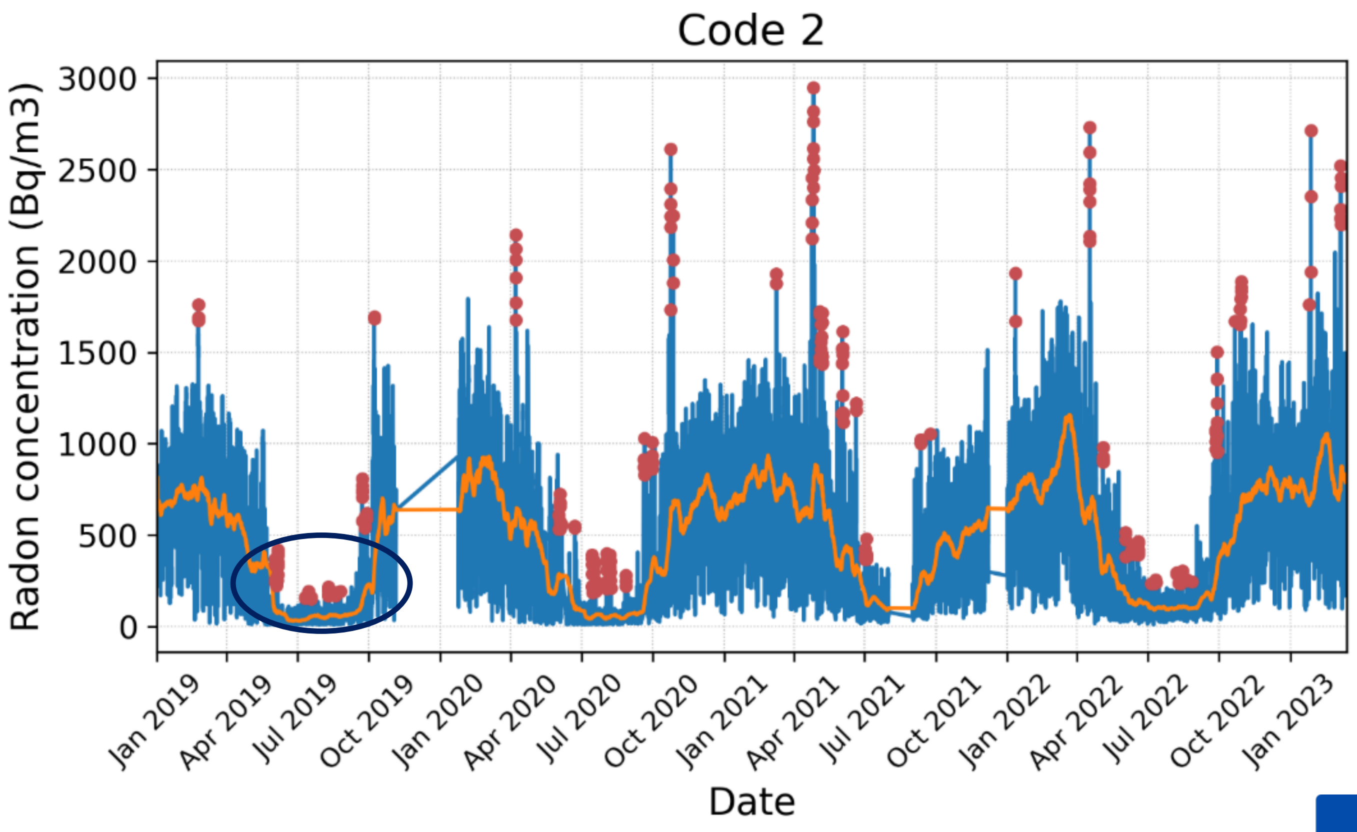


Figure 8. Statistical method (μ+3σ), **with μ și σ calculated for every specific month**, for houses code 5 (left) and house code 2 (right)



- **Monthy threshold obtained for every specific month**

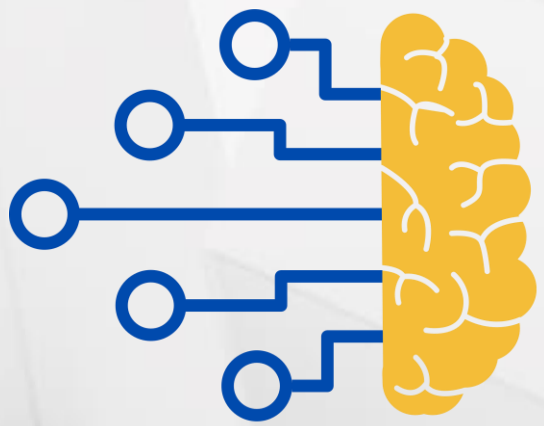
UNSUPERVISED LEARNING METHODS

- Involves training the algorithm with unlabelled data (data for which we do not know the correct answer, we do not know exactly where the anomalies should occur)

- Concentrate on the data learning structure to extract the signal of interest from the noise

Unsupervised methods for determining anomalies include:

- clustering methods (**DBSCAN**, K-means) – meteorological parameters data, spatial data analysis, network analysis etc.
- dimensionality reduction methods (**Autoencoders**, Principal Component Analysis) – image data, text data, geospatial, et
- distance-based methods
- prediction methods
- hybrid methods



Reference:
Chandola, V., Banerjee, A. & Kumar, V. 2009. Anomaly detection: A survey. ACM Comput. Surv. 41, 3, Article 15 (July 2009), 58 pages. https://doi.org/10.1145/1541880.1541882

Density-based clustering algorithm (DBSCAN)

- ❑ Basic parameters:
  - ❑ **ε** - neighbourhood radius
  - ❑ **MinPts** - minimum number of points required to form a cluster

- ❑ Each point looks for other points in the neighbourhood (at distance **ε**) to form a cluster
- ❑ If neighbouring points reach the clustering condition **(MinPts),** that point becomes a **core point**
- ❑ Points that do not belong to any cluster - anomalies, noise

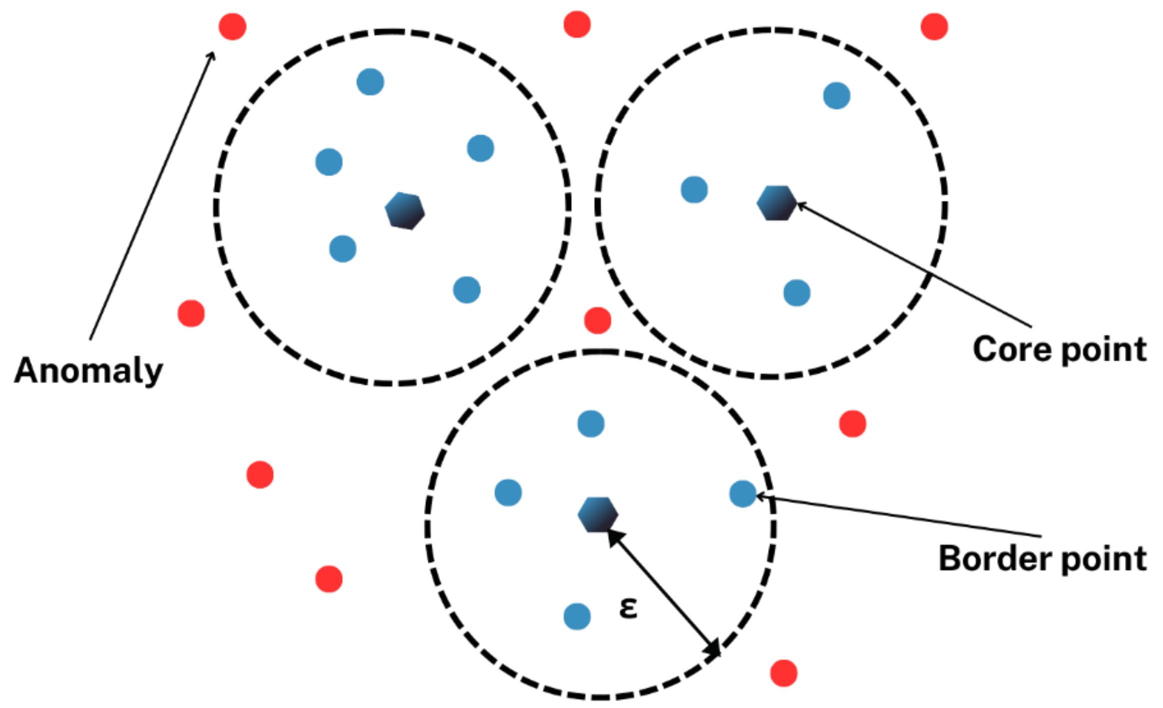


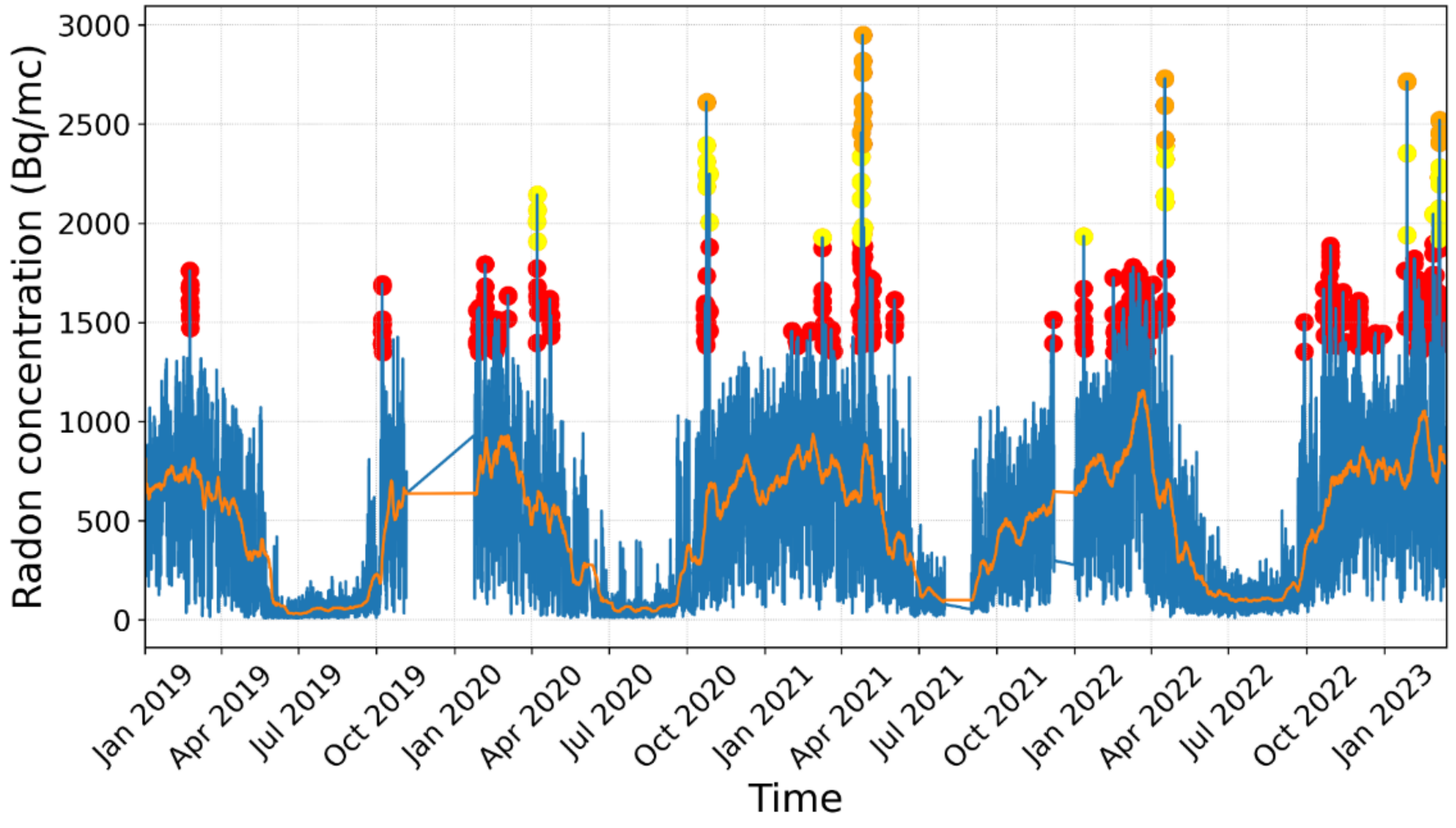
Figure 9. *Density-based clustering algorithm principle (modified after Chauban, 2022)*

Disadvantage: If the dataset has observations that do not have enough neighbours or has anomalies with enough close neighbours, the techniques fail to classify the data correctly, resulting in missed anomalies. Specific disadvantage for clustering methods that use distance between points.

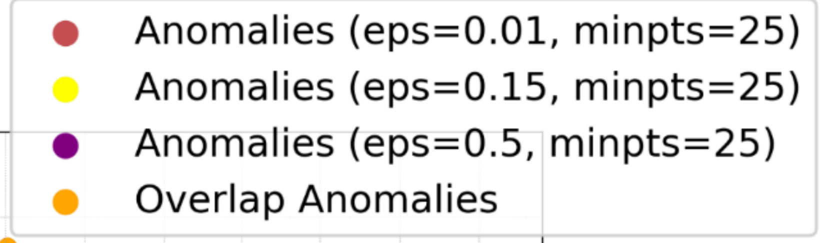
Reference:
Chauhan, N.S. 2022. *An introduction to the DBSCAN algorithm and its implementation in Python.* Machine Learning. https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html

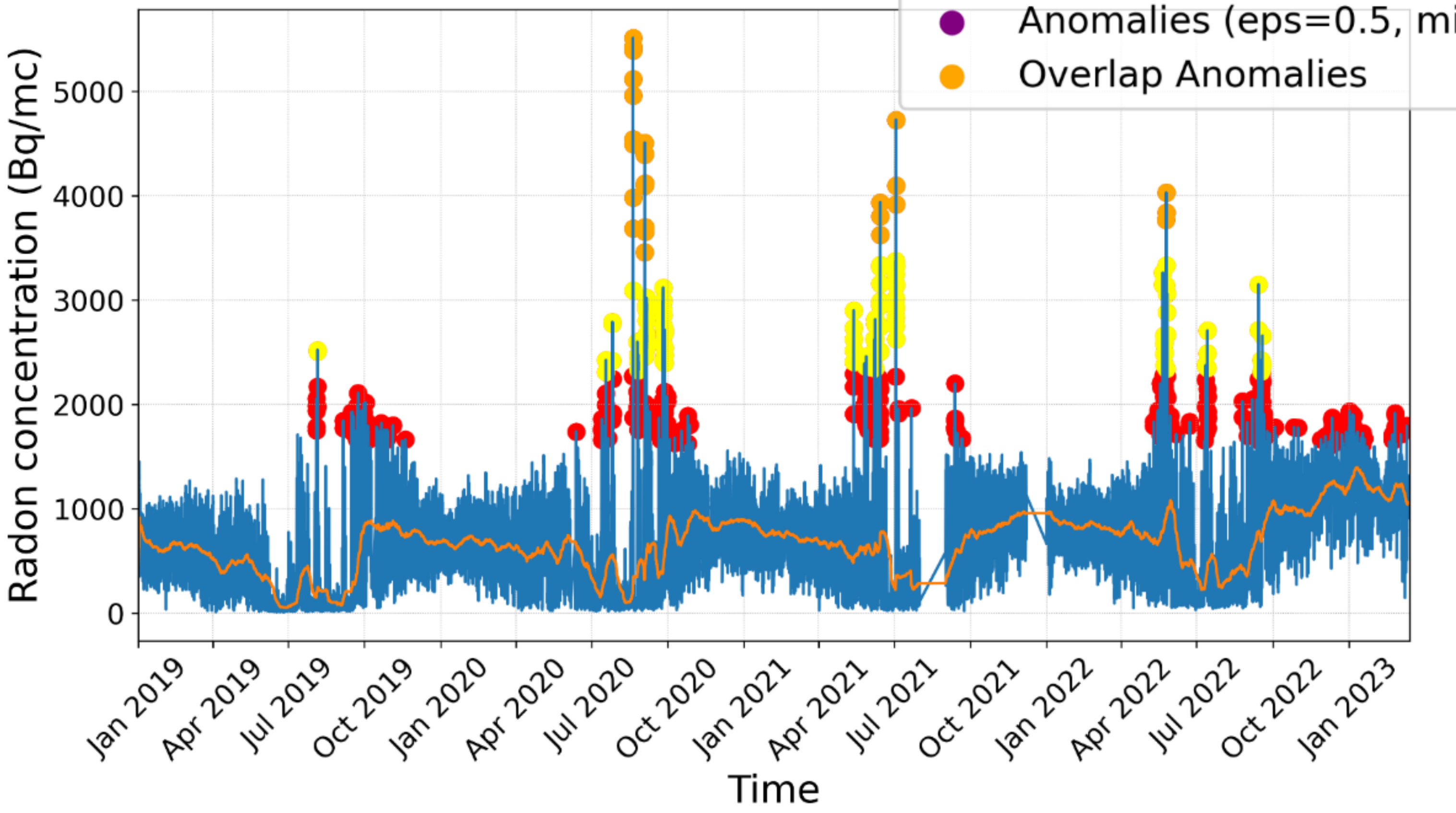
RESULTS

ε values: 0.01, 0.15, 0.5
minpts= 25



3% overlap

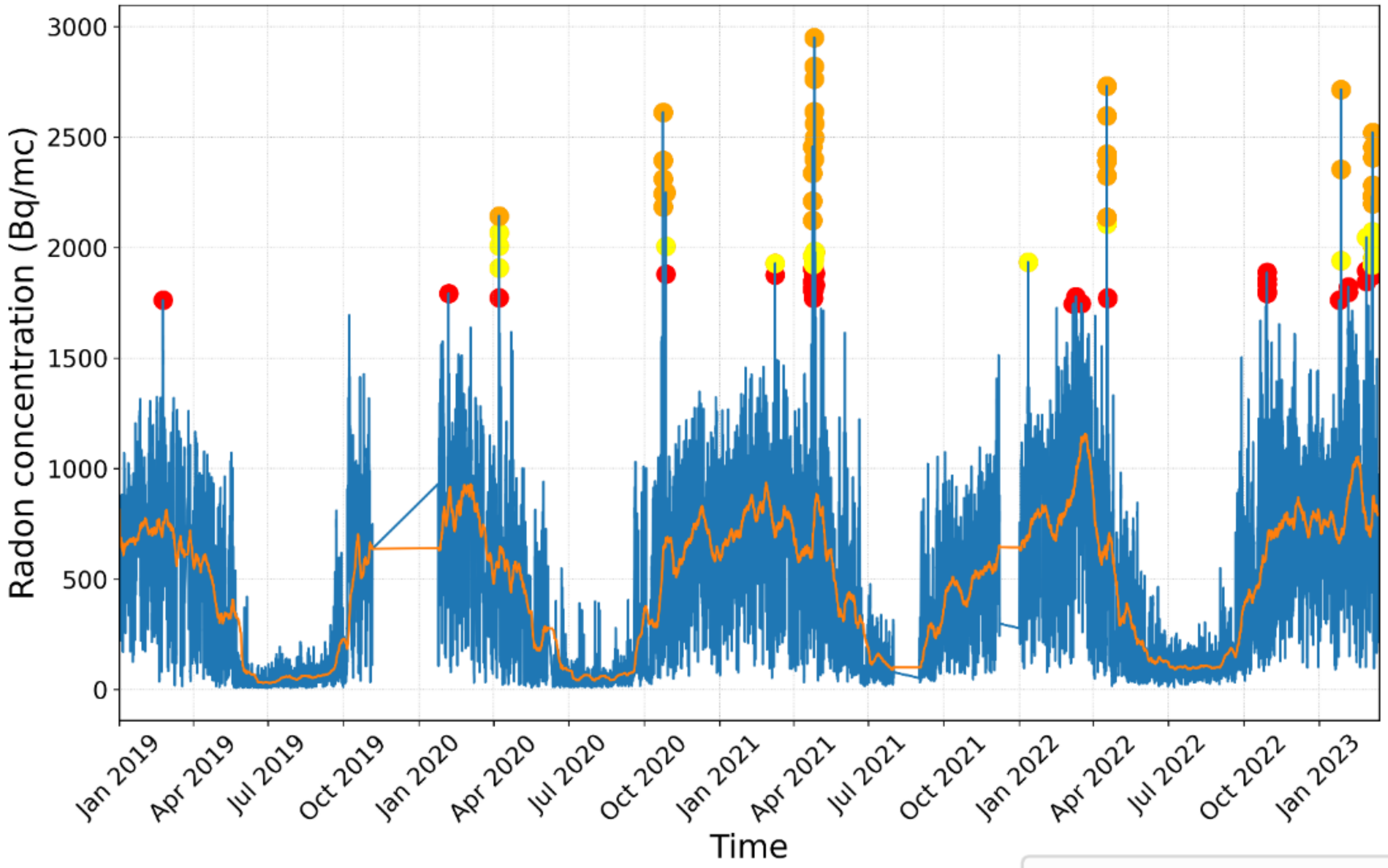




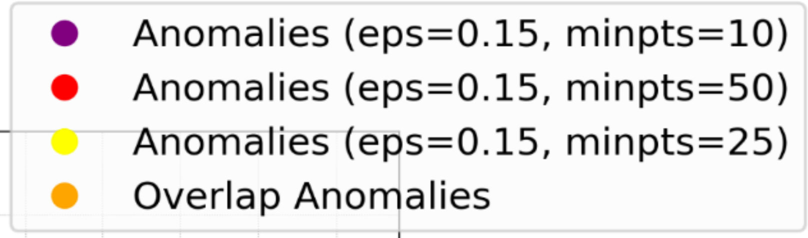
6 % overlap

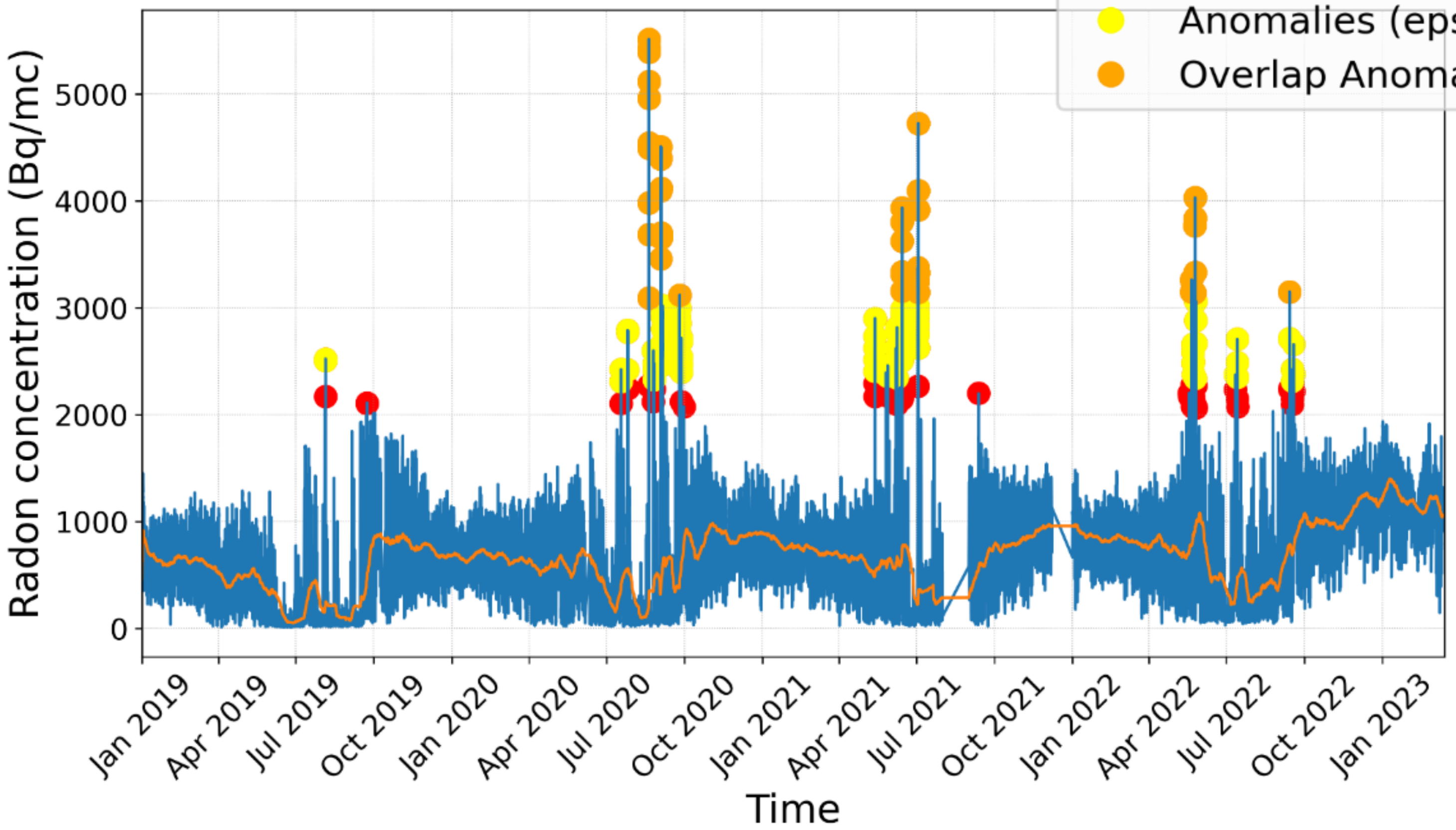
Figure 10. The effects of varying ε parameter in DBSCAN model, house code 2 (top) and code 5 (bottom)

minpts values: 10, 25, 50
ε = 0.15



37% overlap





25% overlap

Figure 11. The effects of varying minpts parameter in DBSCAN model, house code 2 (top) and code 5 (bottom)

RESULTS

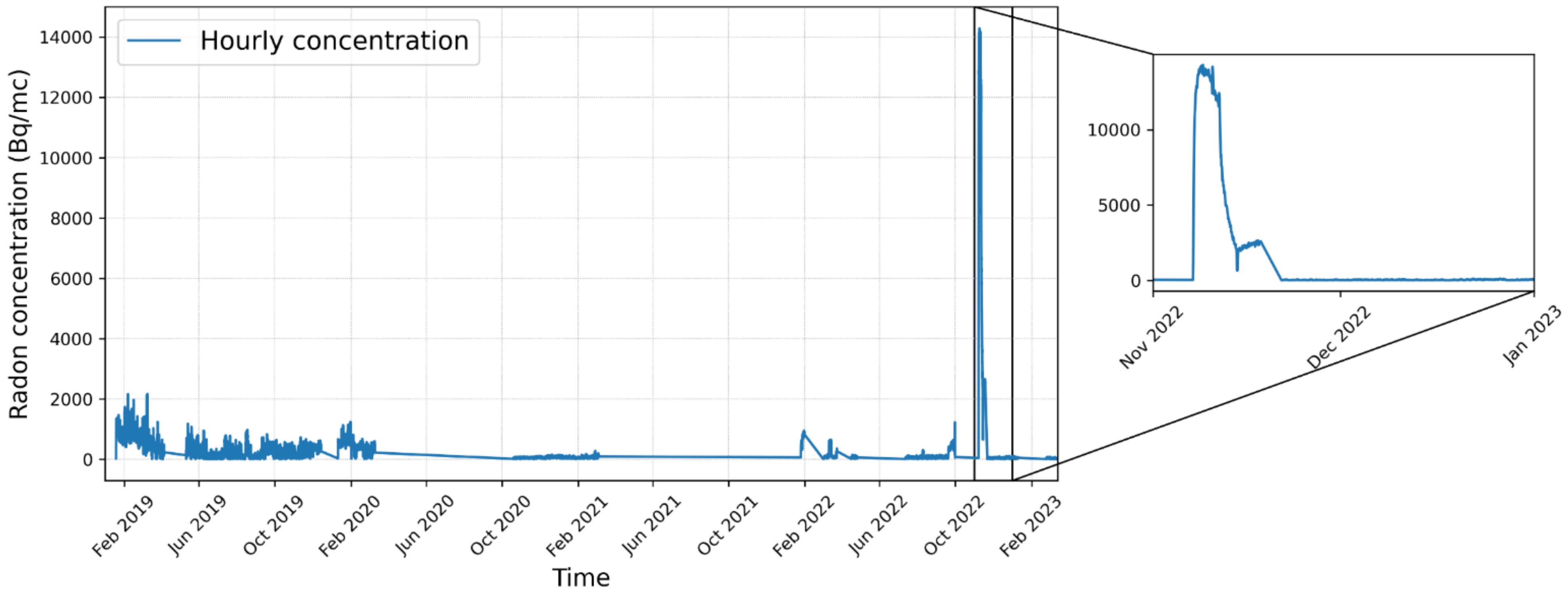


Figure 12. Radon concentration evolution over the period 2019-2024 and zoom in on the period of interest (01.11.2019- 15.03.2023)

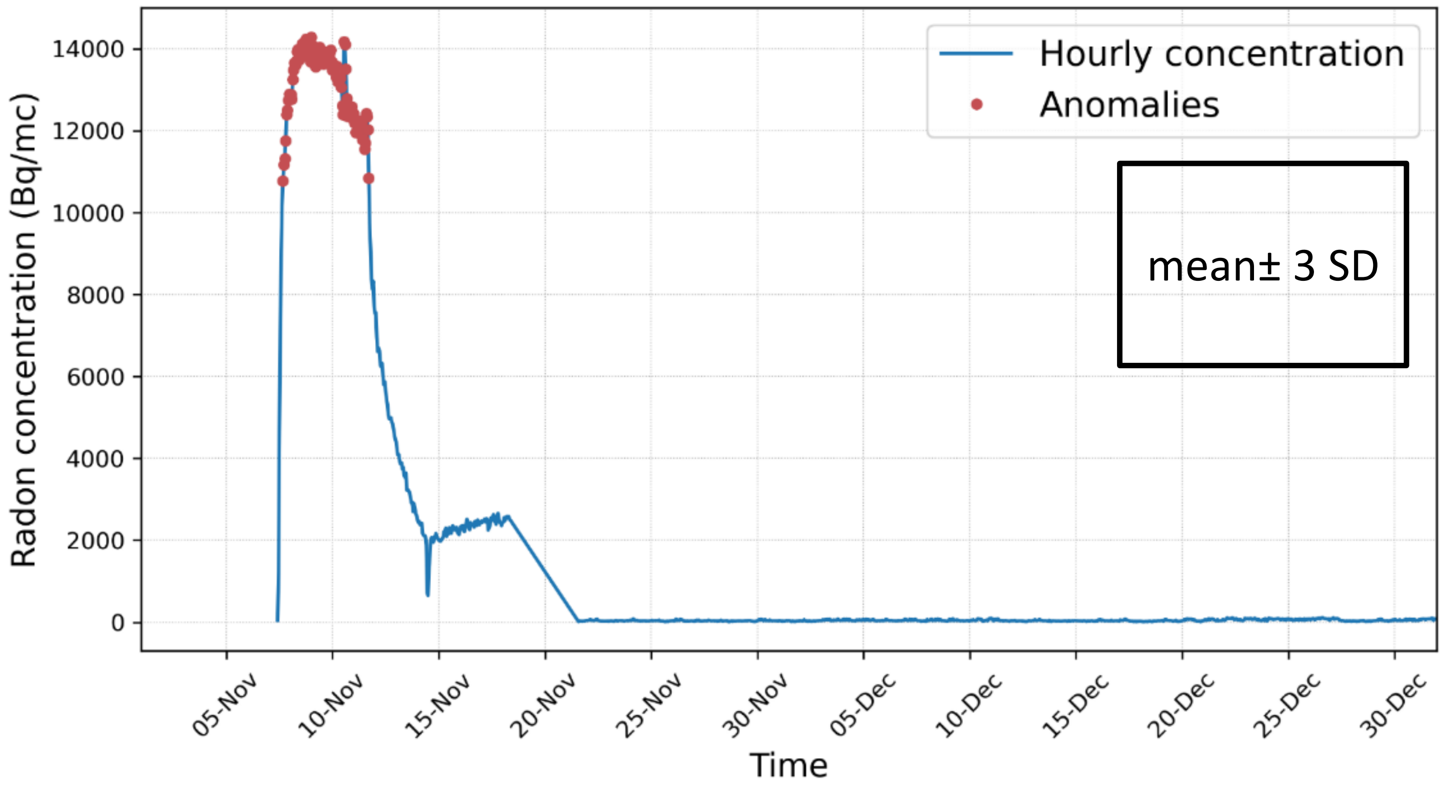


Figure 13. *Application of the statistical method (mean±3 standard deviations)*

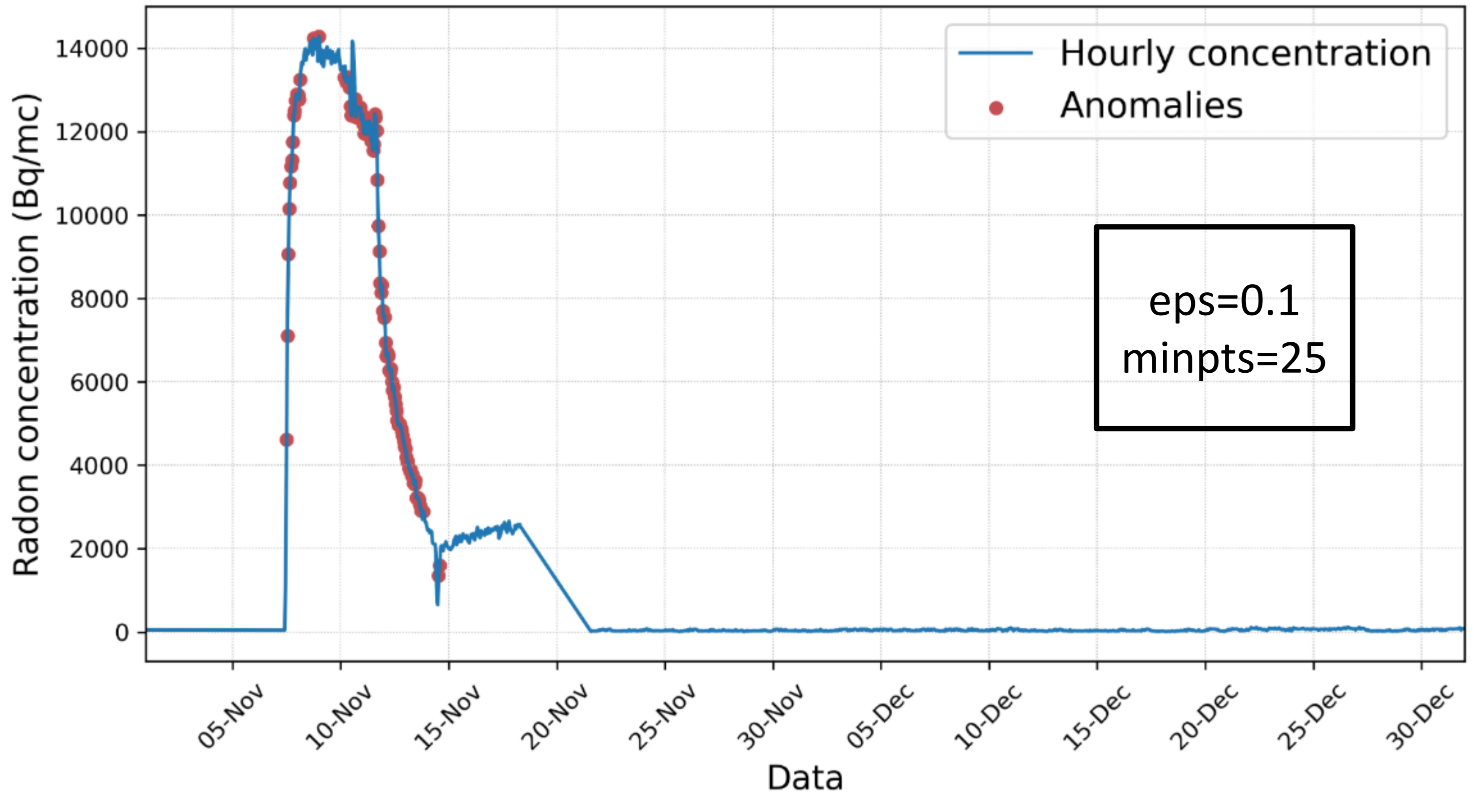


Figure 14. *DBSCAN method application with parameters (eps=1 and minpts= 25)*

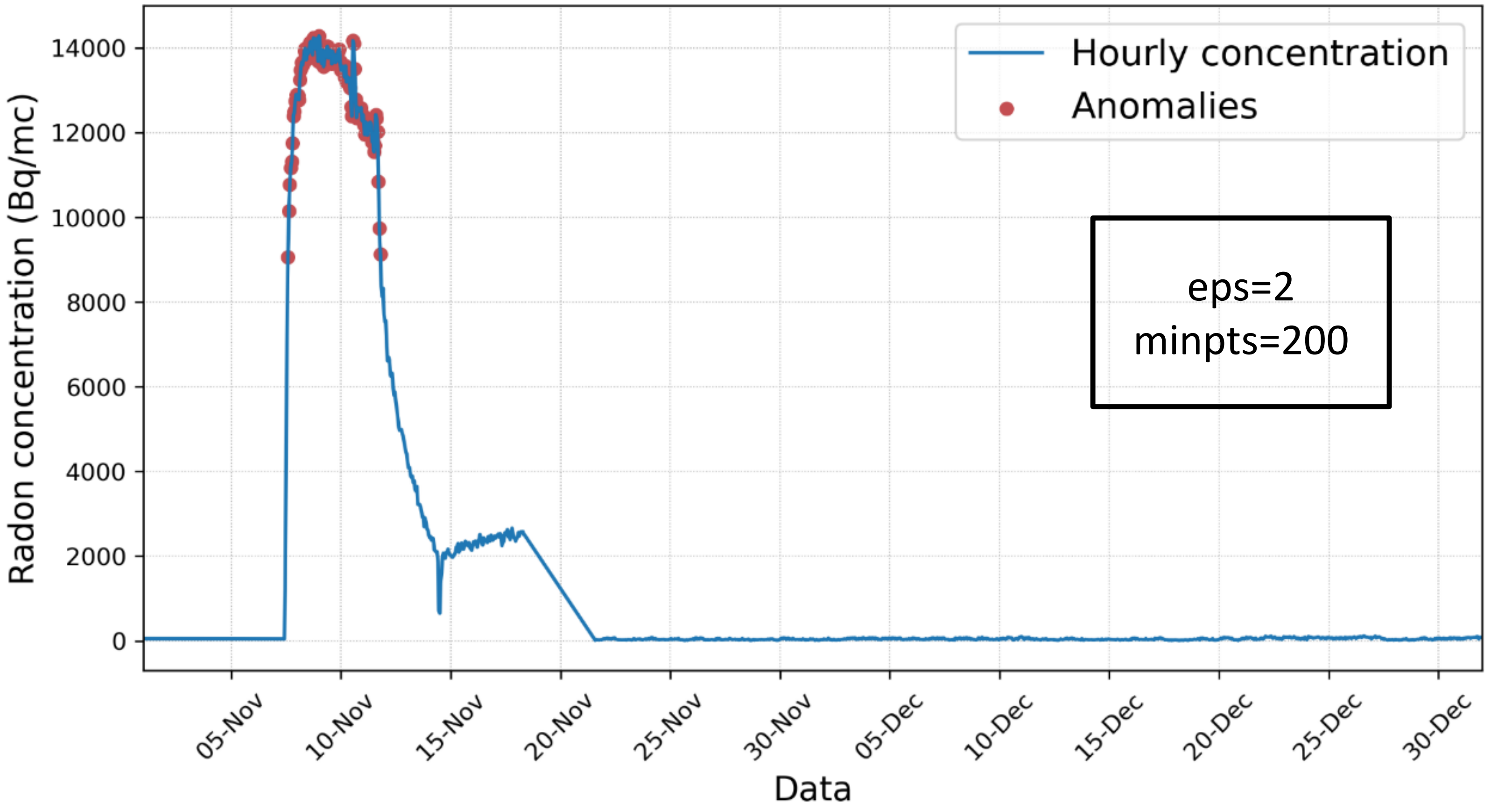


Figure 15. *DBSCAN method application with parameters (eps=2 and minpts= 200)*

Autoencoders

- The autoencoder represents a class of neural network consisting of :
  - **Encoder** - neural network capable of compressing the input into a low dimensionality space, called latent space
  - **Decoder** - is also a neural network, similar in structure to the encoder, but aims to reconstruct (enlarge) the latent space back to the original dimensions of the input

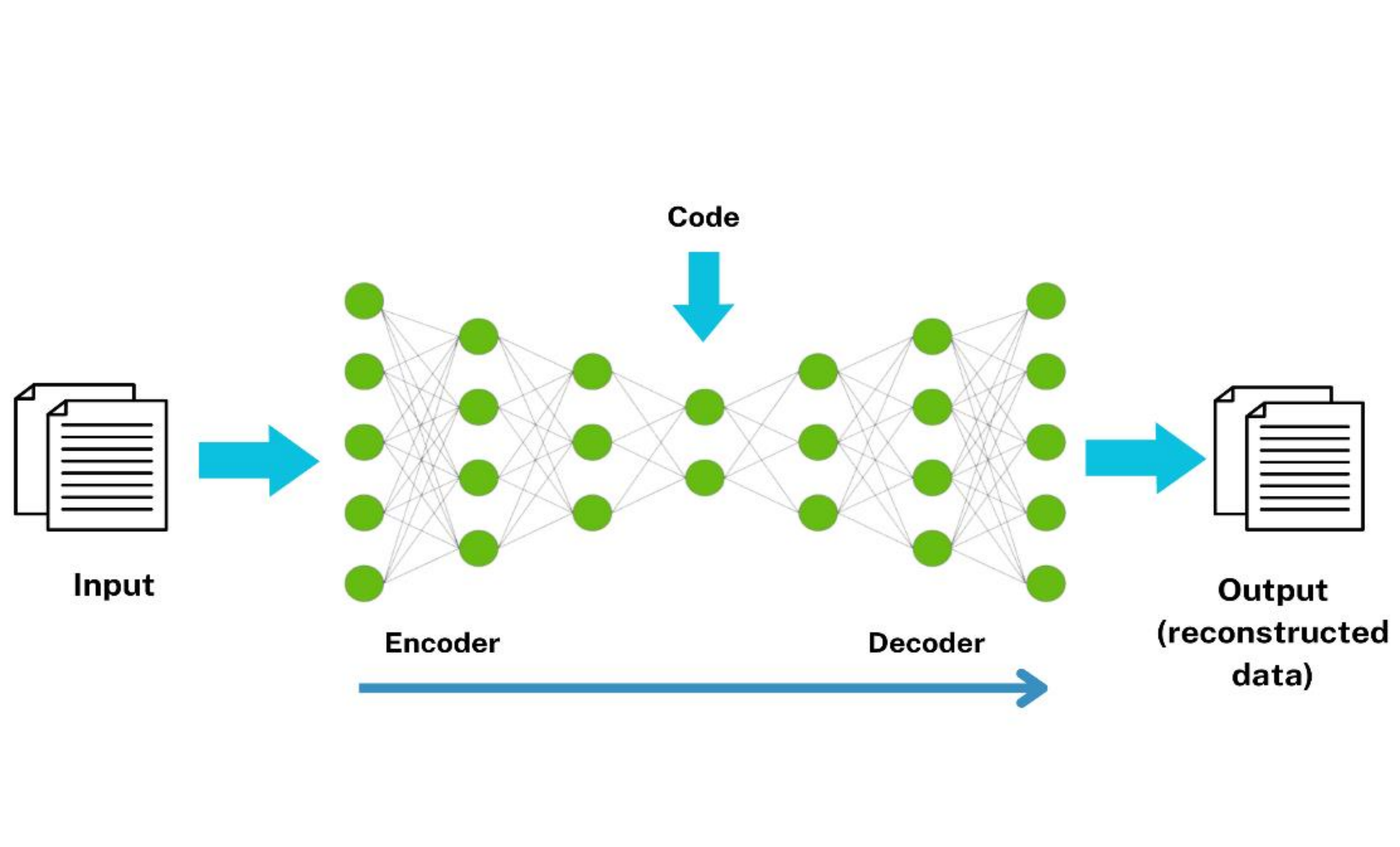


Figure 16. Principle of operation of autoencoder

- Sequential autoencoder model with a Long-Short Term Memory (LSTM) architecture.

- Sequential autoencoders are capable of capturing significant features of time series (sequential data) and detecting anomalies by comparing input data with their reconstruction.

Reference:

Kramer, M. , A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, *37*(2), 233–243

RESULTS

Dataset

Training Set	Test Set
✓ 3 years ✓ 01.01.2019-01.02.2022 ✓ 68%	✓ 1 year ✓ 01.02.2022-13.03.2023 ✓ 32 %

Prediction on test data time period → **Reconstruction error - MAE** → **Set the maxMAE threshold** → ANOMALIES

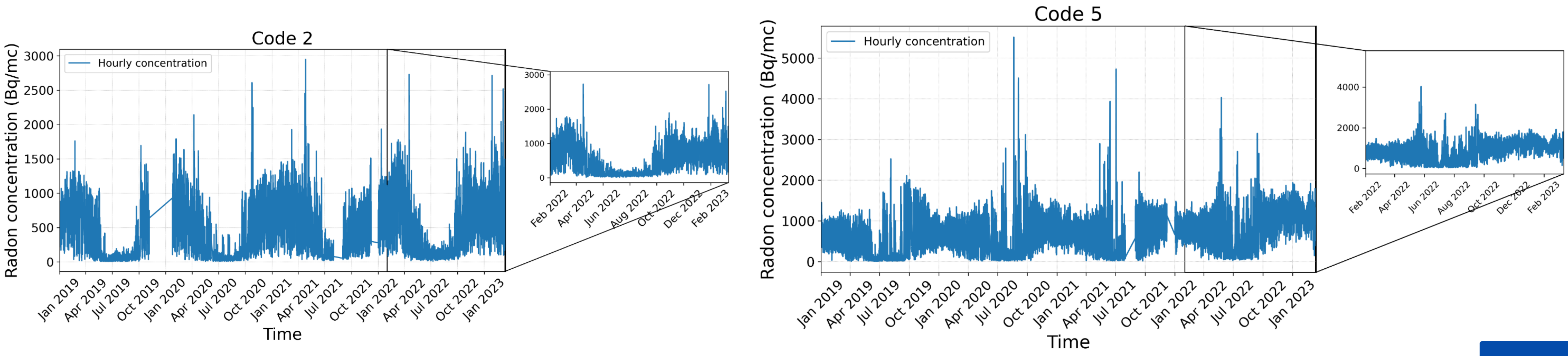


Figure 17. *Radon concentration time series and training/validation data selection for house code 2 (left) and code 5 (right)*

RESULTS

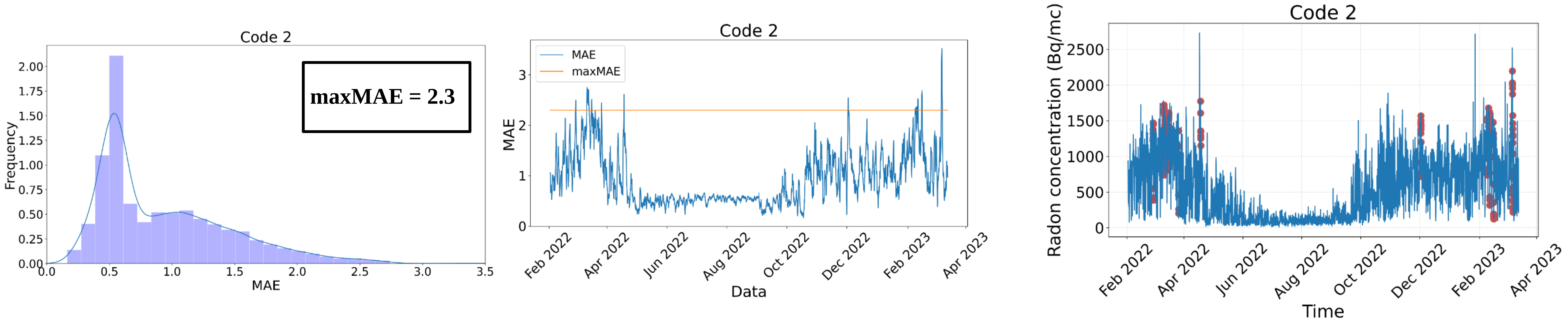


Figure. 18 *MAE frequency histogram (left), maxMAE threshold detection graph (middle) and anomaly detection for radon concentration using autoencoders (right), for the investigated time period, for house code 2*

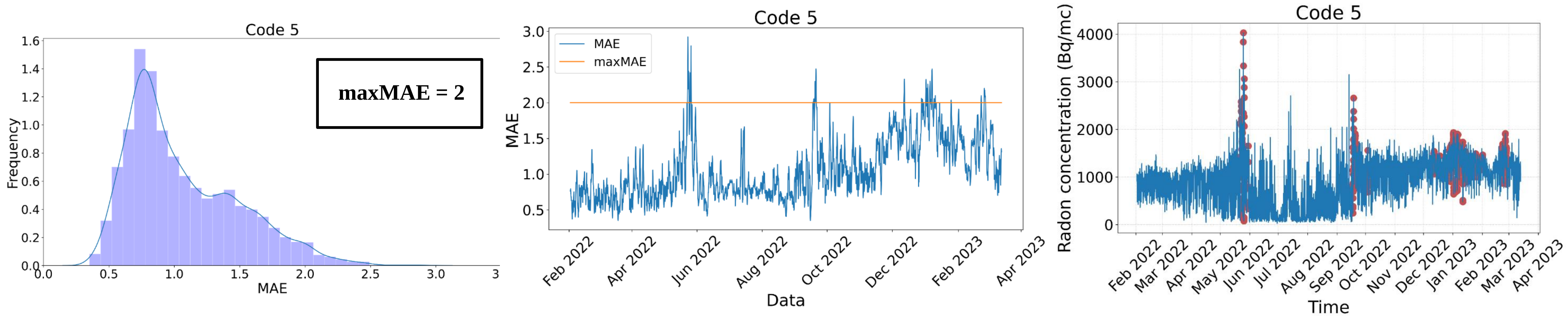


Figure 19. *MAE frequency histogram (a), maxMAE threshold detection graph (b) and anomaly detection for radon concentration using autoencoders (c), for the investigated time period, for house code 5*

RESULTS

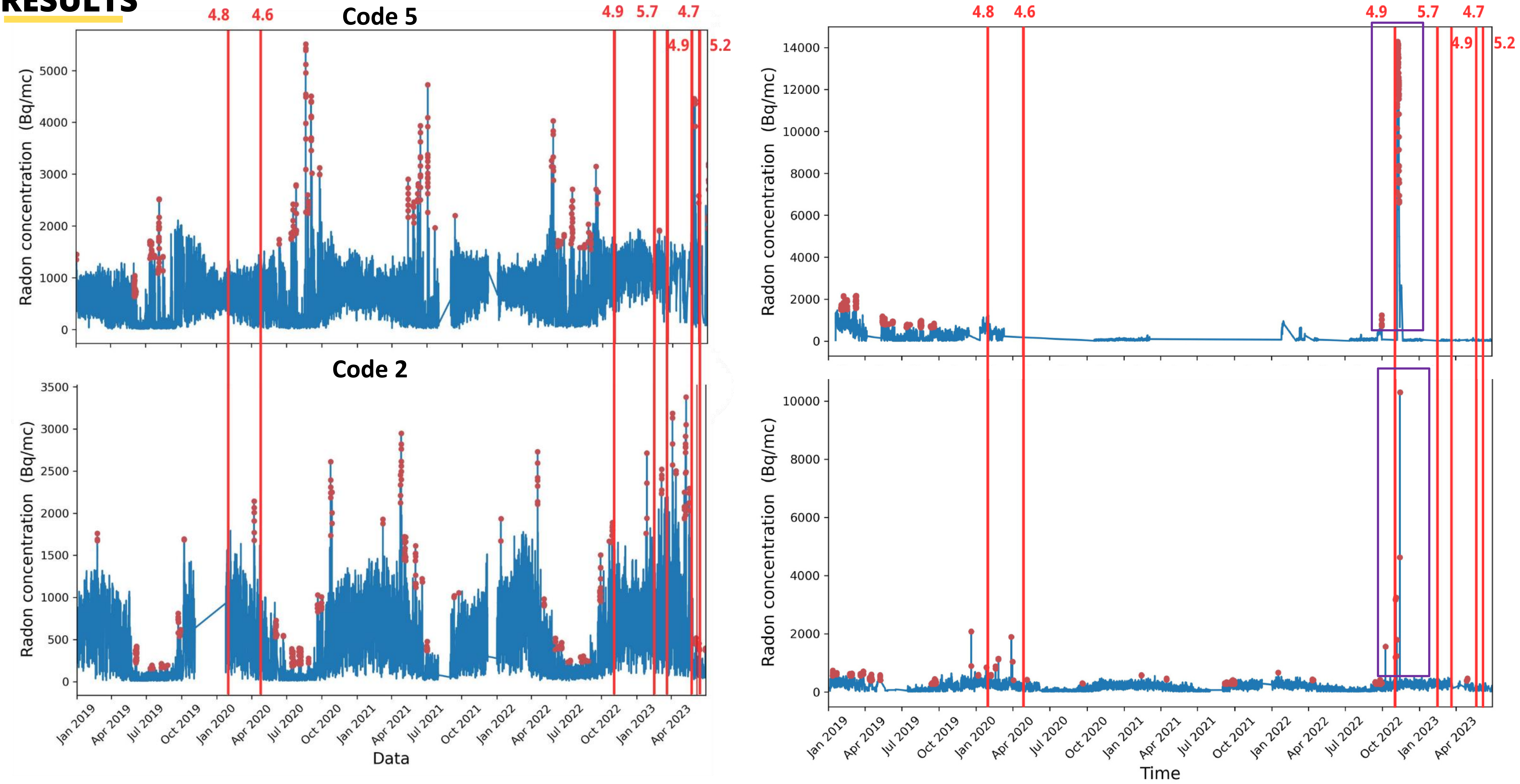


Figure 20. *Radon concentration evolution for 4 houses, with determination of anomalies by statistical method, with parameters calculated at seasonal level and highlighting earthquakes with Mw>4.5*

RESULTS

DBSCAN
ε=0.15
minpts=50

Statistical method - seasonal
Threshold of 3σ

Statistical method - monthly
Threshold of 3σ

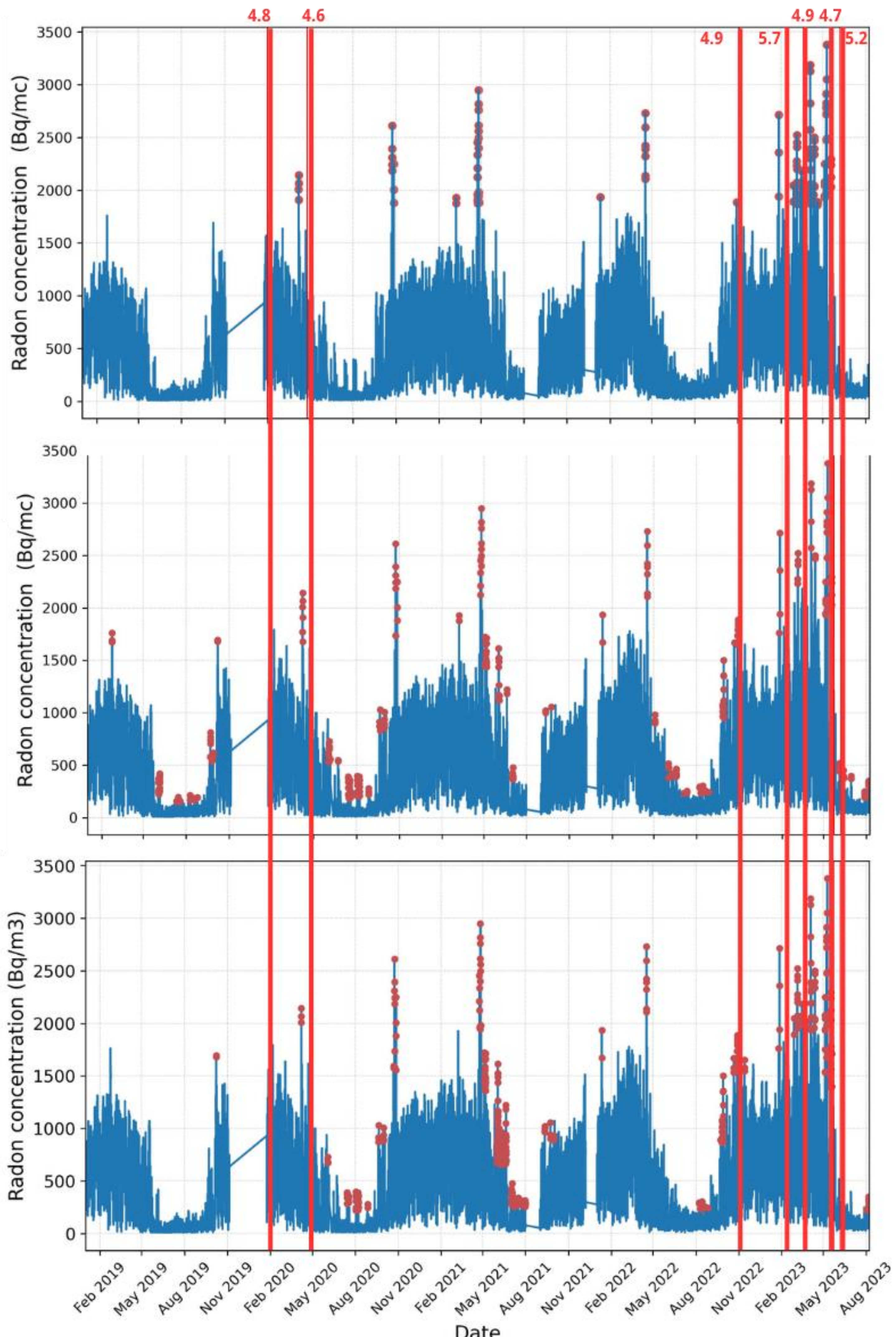
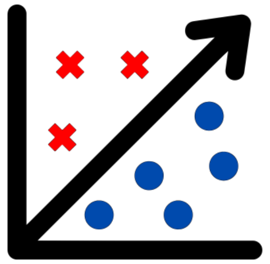


Figure. 21 *Aplication of the DBSCAN method (top), statistical method with seasonal calculated parameters (middle) and statistical method with monthly parameters (bottom)*

Conclusion and perspectives

- Application of anomaly determination techniques on the data set
- Technical advantages/disadvantages
- Selection of a technique → specific to the purpose of the study



Perspectives:

- Applying other techniques and observing their peculiarities.
- Testing supervised learning methods (with labelled data) - Neural networks, decision trees, etc.
- Explaining anomalies (seismic events, weather factors, etc.)

THANK YOU!