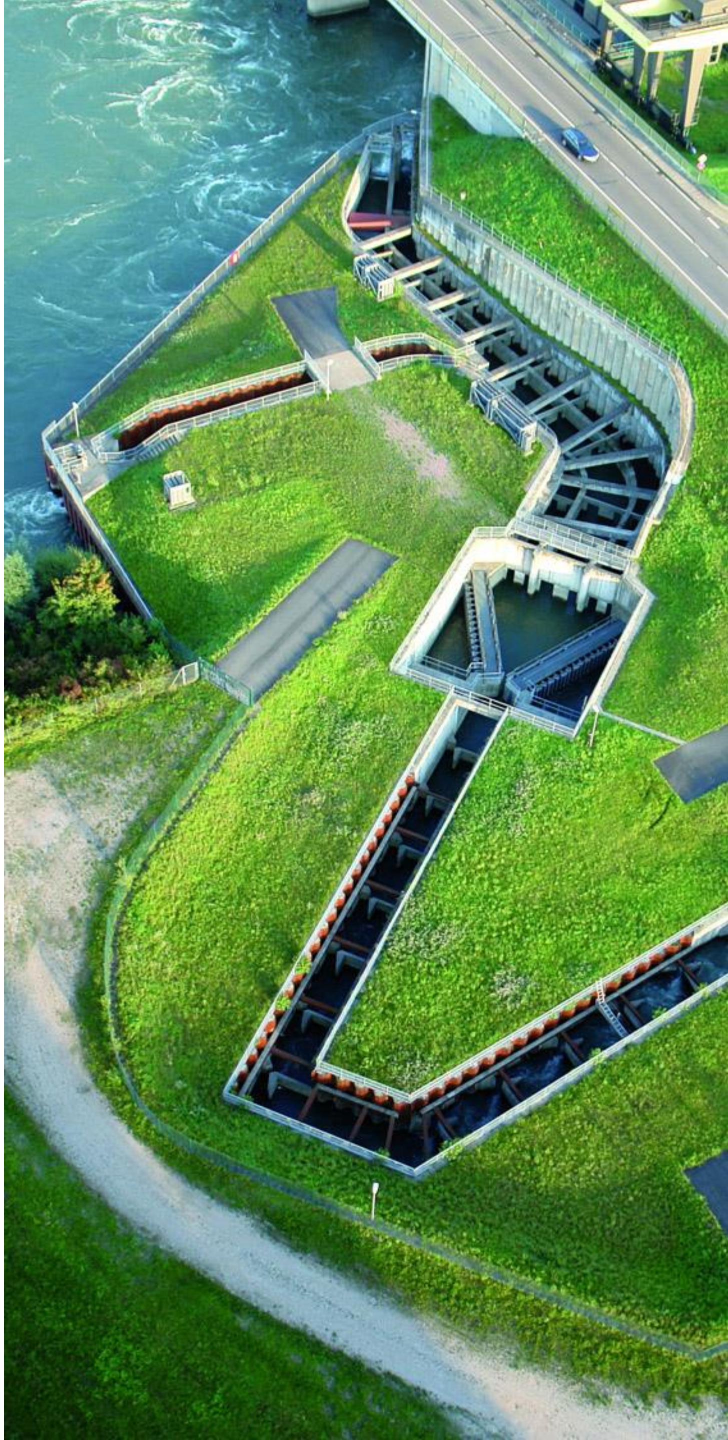
EDF R&D

Détection d'anomalies dans des series temporelles : une approche non paramétrique

Christian Derquenne



Séminaire de statistique appliquée du CNAM – 28 janvier 2022

Plan

Context and goal

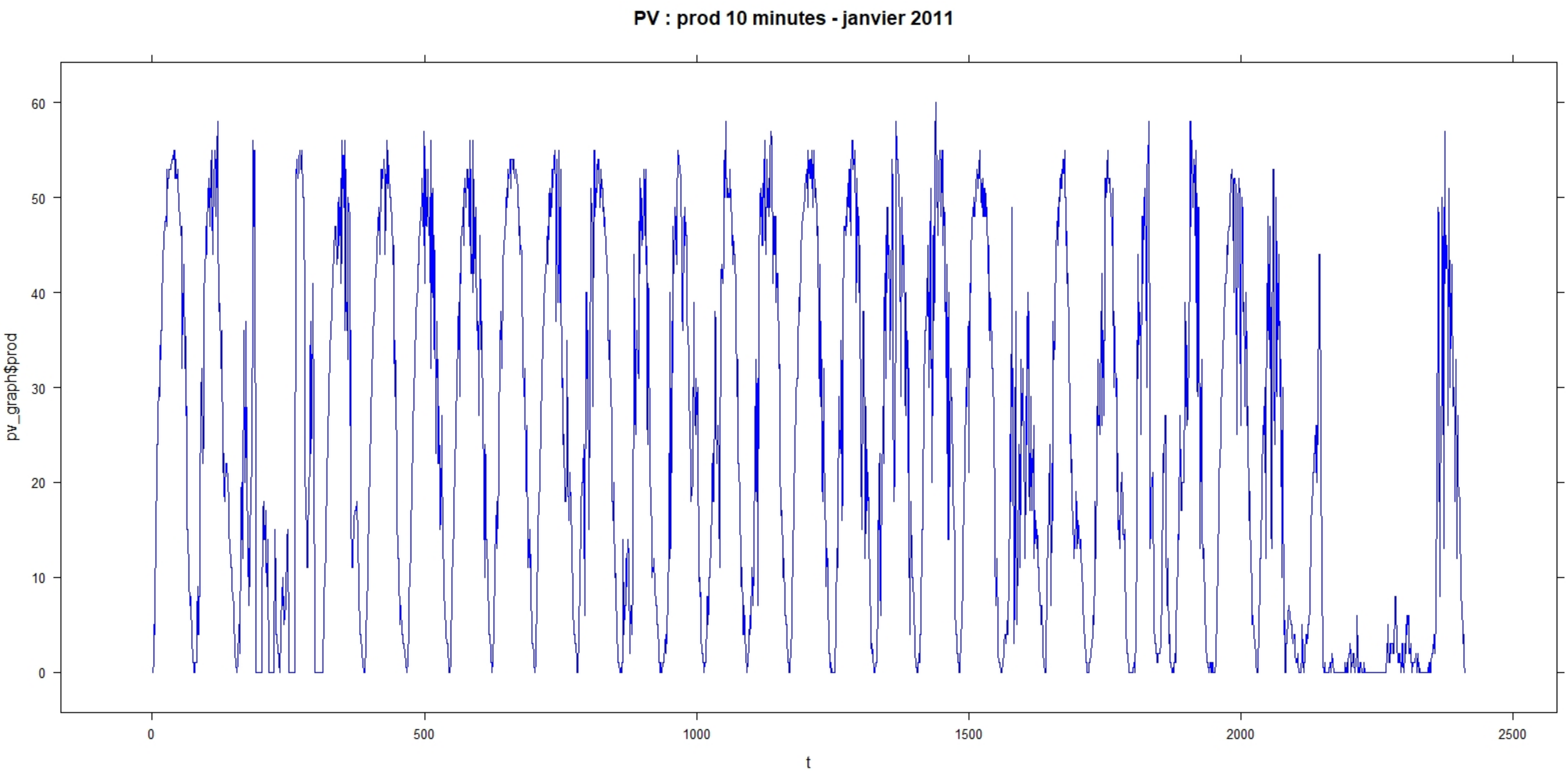
A non parametric approach for anomalies detection

Application of approach on PV data

Contributions, applications and further researches

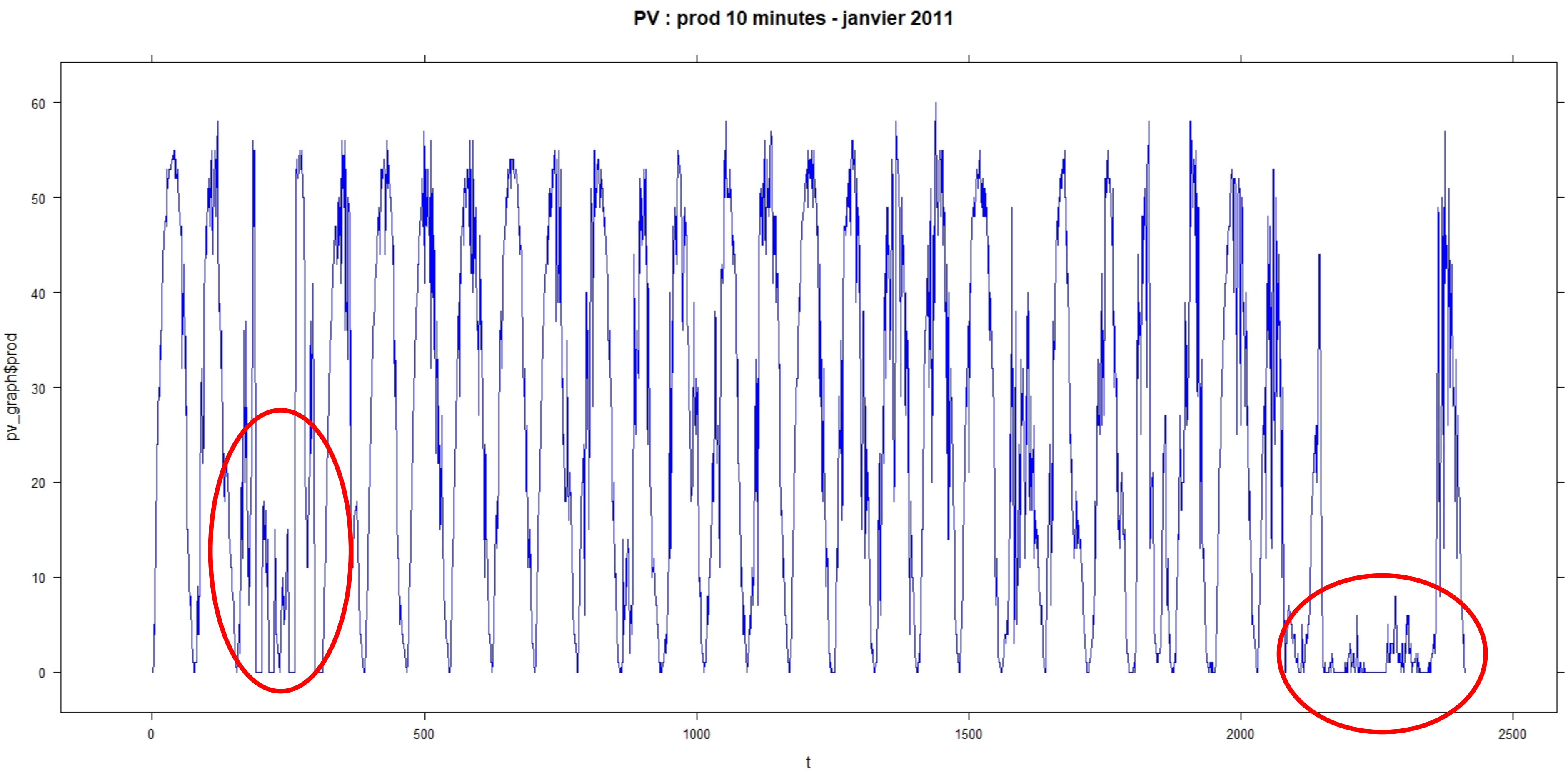
Context and goal

Photovoltaic data (PV stations)



Context and goal

Photovoltaic data : *anomalies detection*



Context and goal

Applications

- ✓ **Using these data in energy management for production or consumption forcasting**
- ✓ **Problem :** quality of results can be affected in presence of outliers, anomalies, breaks, …
- ✓ **Solutions :** detecting anomalies, breaks before modelling; robust model approaches

Some definitions

- ✓ **Breaks detection** = Search for changes in behavior (which may be long) in average, in variance, both, in distribution …
  - ➢ **Offline *vs* online**
- ✓ **Anomalies detection** = Search for sudden changes in behavior (of medium or short duration): peaks, troughs, …
  - ➢ **Offline *vs* online**

Context and goal

Anomalies detection : *some approaches*

✓ **Rather non-parametric approaches**

- using robust statistics based on ranks [Friend et al., 2007]
- forcasting models based on LSTM (Long-Short Term Memory) joint to Deep Learning [Maya et al., 2019]
- detection of anomalies on aggregated times series [Derquenne, 2015]
- tests non paramétriques conjoints à une approche bayesienne (modèle Bernoulli Detector) pour séries temporelles univariées et multivariées [Harlé, 2006]
- utilisation de réseaux bayesiens (graphes de dépendance entre signaux) [Harlé, 2006]

Plan

Context and goal

A non parametric approach for anomalies detection

Application of approach on PV data

Contributions, applications and further researches

Anomalies detection approach

le cnam

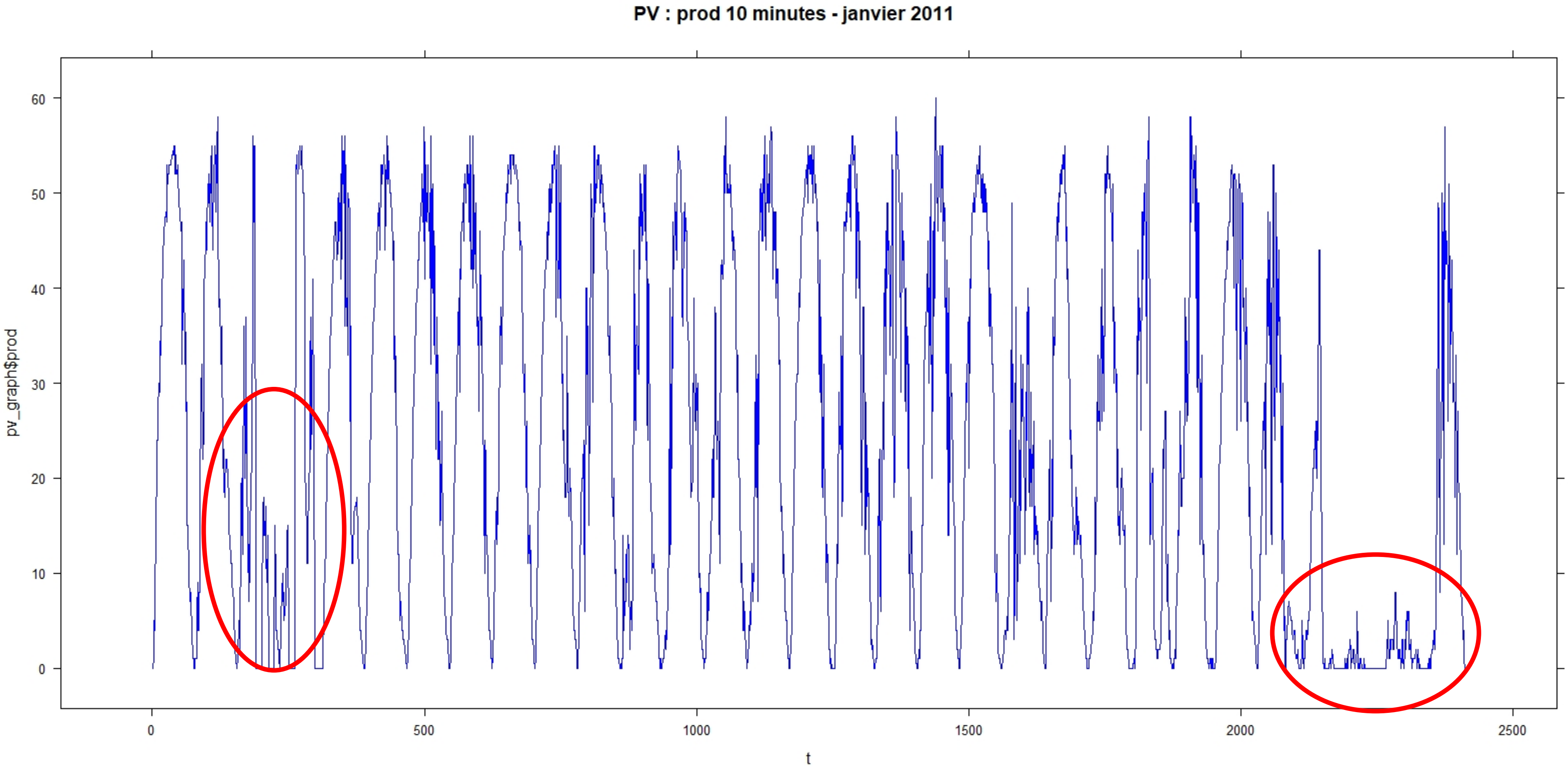
Overall principle of approach

- ✓ **Splitting of time series**
- ✓ **Categorization of time series**
- ✓ **Aggregation of data** (for instance : 10 min → 1 hour)
- ✓ **Building of aggregating marginal distributions of the categories**
- ✓ **Building of the statistic of test**
- ✓ **Building of empirical distribution empirique of test**
- ✓ **p-value calculation**

Anomalies detection approach

The approach step by step (1)

✓ **The data**



Anomalies detection approach

The approach step by step (2)

Let's $Y_{t,(t=1,T)}$ be a seasonal regular time series

→ For instance, Y_t is a PV production extracted by 10 minutes

→ T = number of observed data by 10 minutes

✓ **Splitting of time series in *M* categories**

→ allows to reason of comparable quantities (for ex, by month)

$$I_0 = 0 \qquad I_1 = \left[0\,;\max_t Y_t / M\right]$$

$$I_m = \left[(m-1)\times\max_t Y_t / M; m\max_t Y_t / M\right]$$

...

$$I_M = \left[(M-1)\times\max_t Y_t / M;\max_t Y_t\right]$$

...

where m is the number of the category

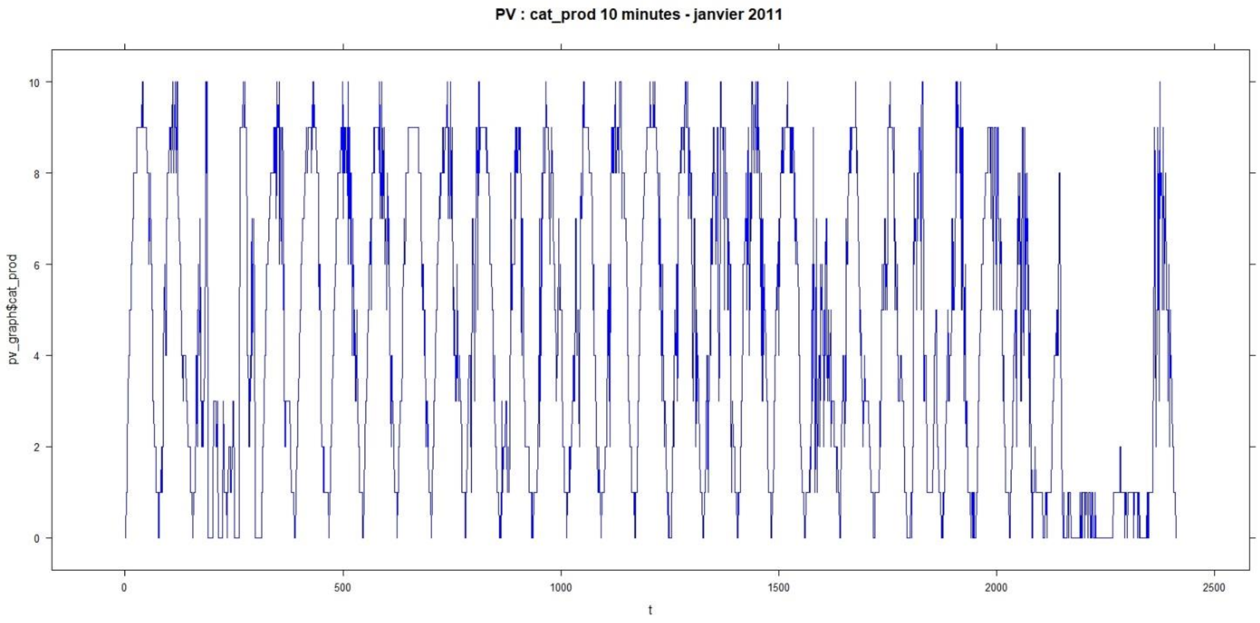
Anomalies detection approach

The approach step by step (3)

Categorisation of the time series

→ comes from of previous splitting and allows to reason on ordinal categorical values

$$b_t = m \times 1_{[Y_t \in I_m]} \quad \text{where } m \text{ is the number of category}$$



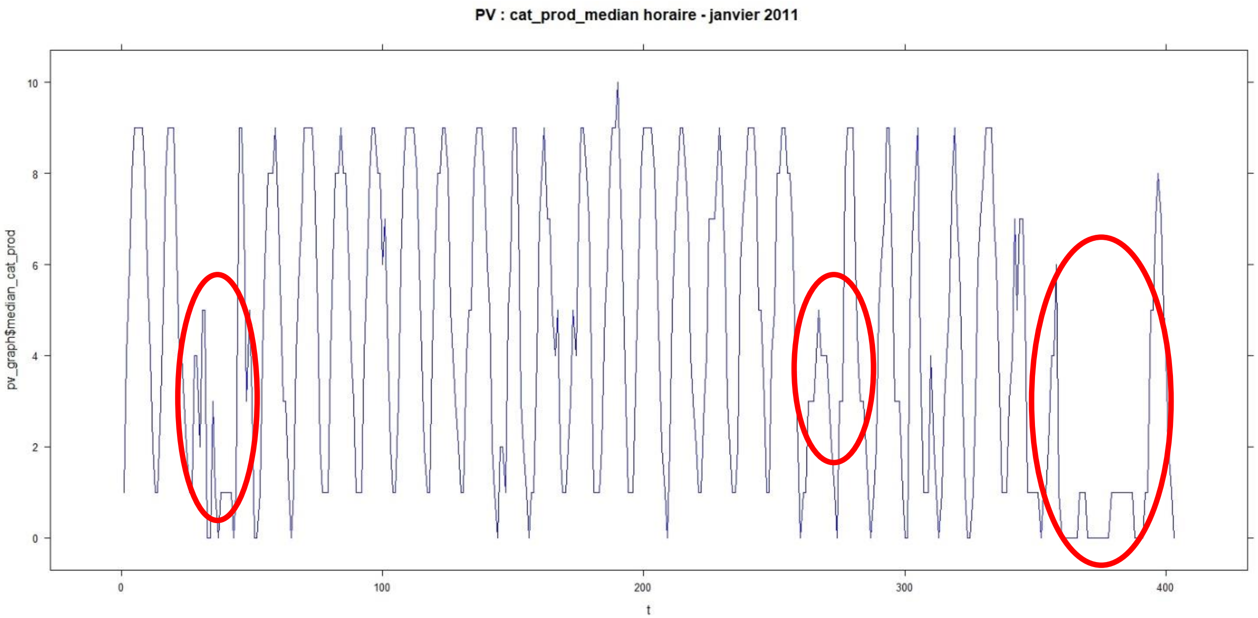
Anomalies detection approach

The approach step by step (4)

Aggregation of data

→ allows to summarize, if necessary, categorical data (10 min to 1 hour)

$$b_{j,h} = \operatorname*{\mathbf{med}}_{t \in (j,h)} b_t \quad \text{where } h \text{ and } j \text{ are respectively : hour } h \text{ of day } j$$



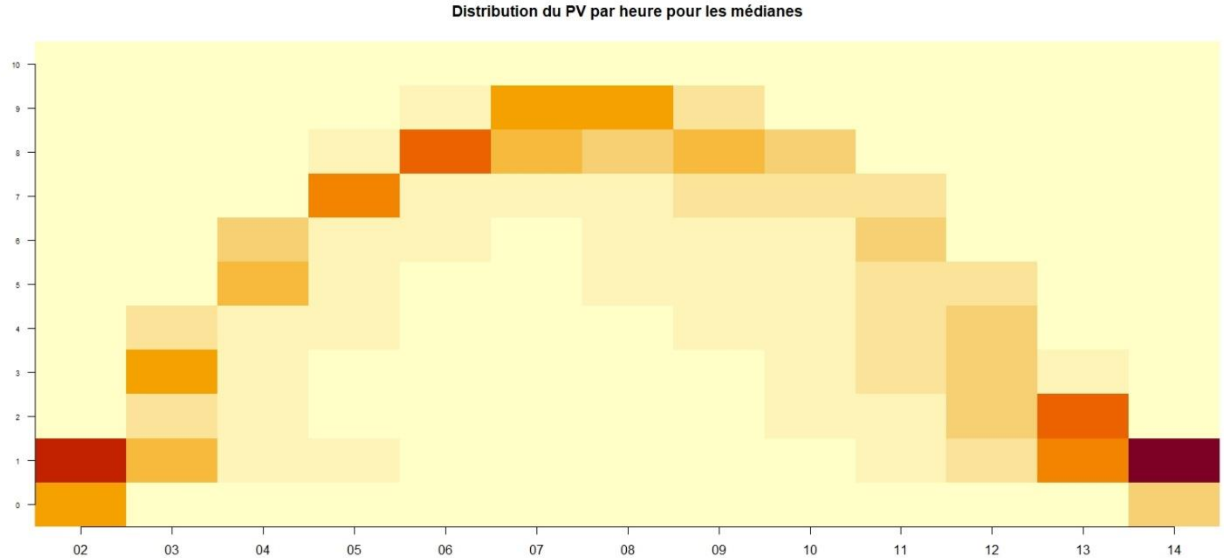
Anomalies detection approach

The approach step by step (5)

Hourly marginal distributions of the categories

→ allows to obtain observed proportions of the categories for each hour

$$f_m(h) = N_m(h) \Big/ \sum_{l=0}^{M} N_l(h) \qquad \text{where} \qquad N_m(h) = \sum_{j=1}^{n_h} 1_{[b_{j,h}=m]}$$



and n_h is the number of days for the study period for the hour h

Anomalies detection approach

The approach step by step (6)

Building of the statistics of test

→ allows to provide the structure of anomalies

$$S_j^{(moy)} = \frac{1}{card(\Omega)} \sum_{h\in\Omega} \sum_{m=0}^{M} 1_{[b_{j,h}=m]} f_m(h) \quad \text{where } j \text{ is the number of day}$$

$$S_j^{(med)} = \underset{h}{\mathbf{med}} \sum_{m=0}^{M} 1_{[b_{j,h}=m]} f_m(h) \qquad S_j^{(LMS)} = \underset{h}{\mathbf{lms}} \sum_{m=0}^{M} 1_{[b_{j,h}=m]} f_m(h)$$

For a day j and an hour h :

$$S_{j,h} = \sum_{m=0}^{M} 1_{[b_{j,h}=m]} f_m(h)$$

Interpretation : The lower the value of the statistic for an observation $S_{j,h}$ the lower its frequency of occurrence, the more likely it is to be an "anomaly"

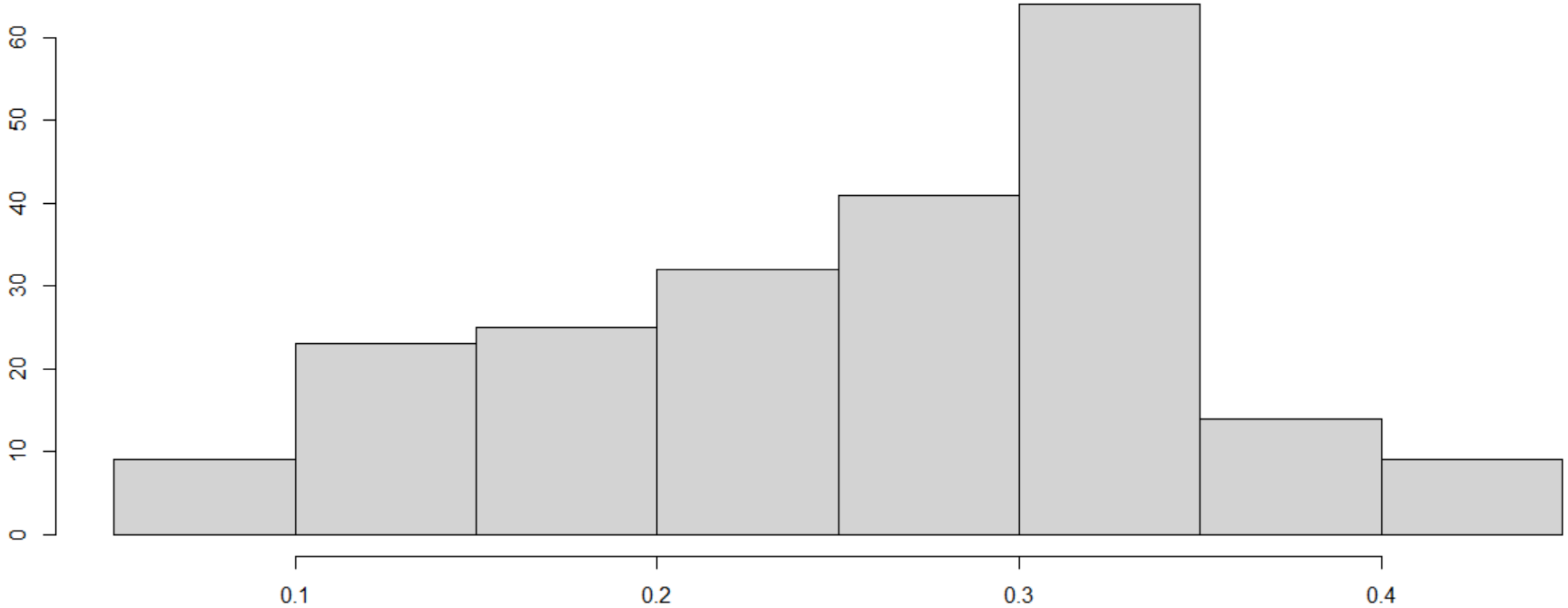
Anomalies detection approach

The approach step by step (7)

Building of empirical distribution of test

→ allows to take into account all the values observed for each day (global) or each hour (individual)

$$F_h(S_{j,h}) = \Pr\lfloor D_h < S_{j,h} \rfloor$$ is the empirique cumulative distribution fonction D_h for $S_{j,h}$



Anomalies detection approach

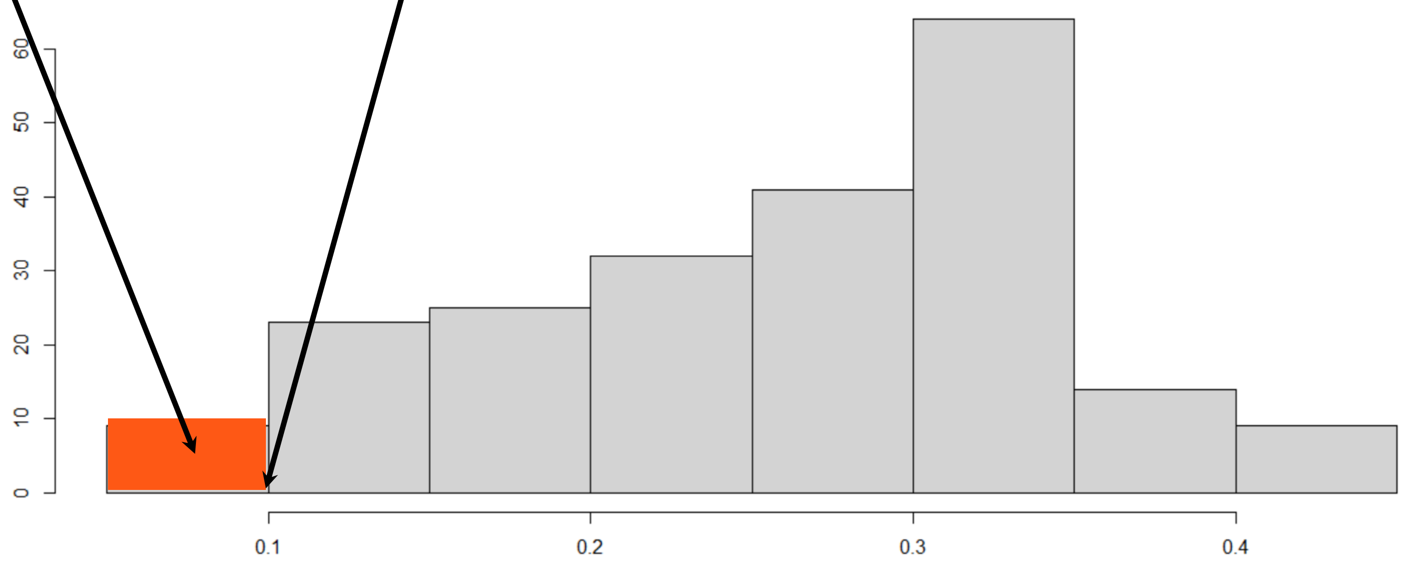
The approach step by step (8)

p-value calculation

→ building under null hypothesis (absence of anomaly) and coming from MC simulations

$\mathrm{H}_0 : Y_t$ is not an anomaly vs $\mathrm{H}_1 : Y_t$ is an anomaly

$$p-val = \Pr\left[D_h < s_{j,h}^{(obs)}\right]$$



Plan

Context and goal

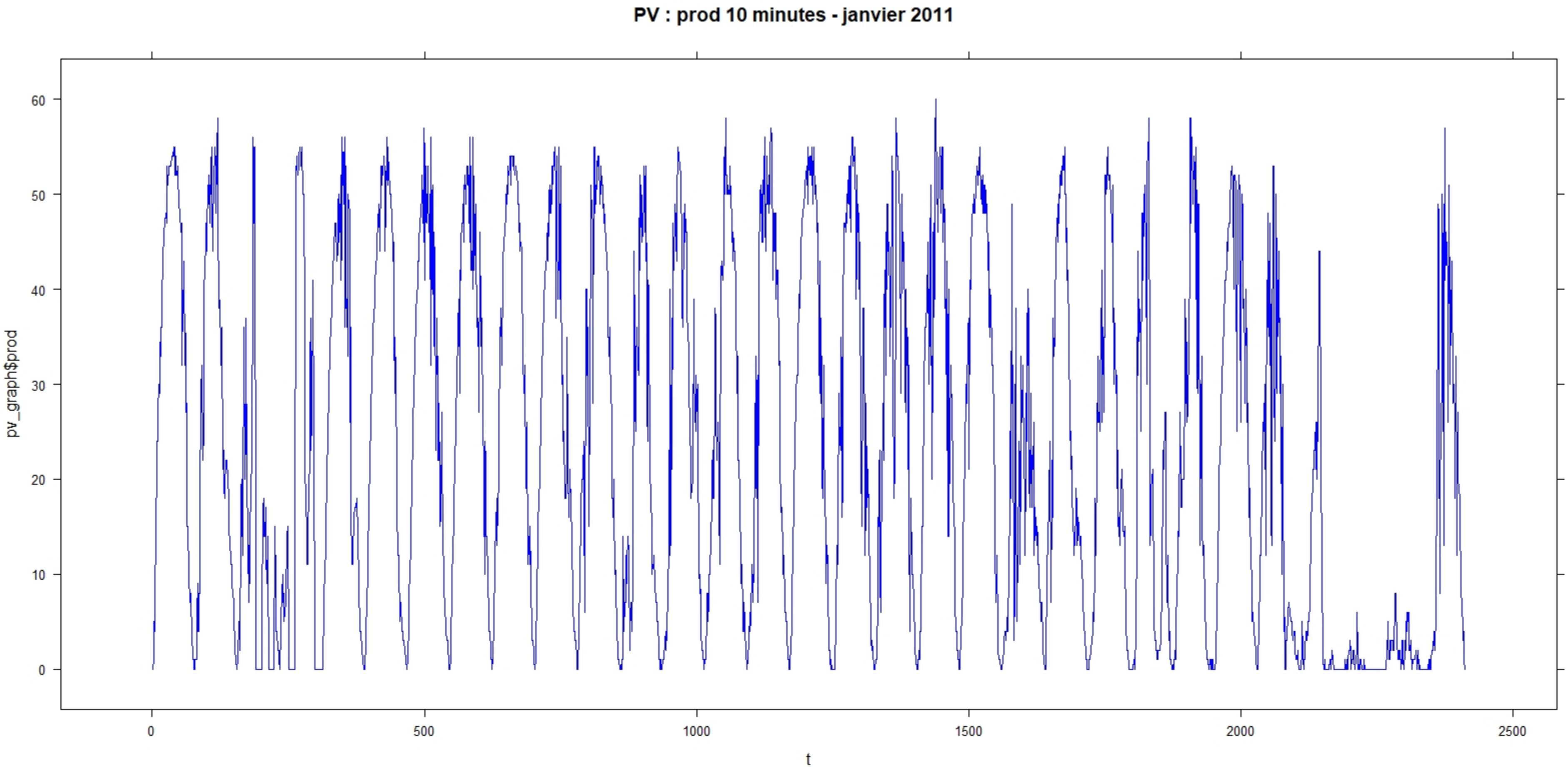
A non parametric approach for anomalies detection

Application of approach on PV data

Contributions, applications and further researches

Application of approach on PV data

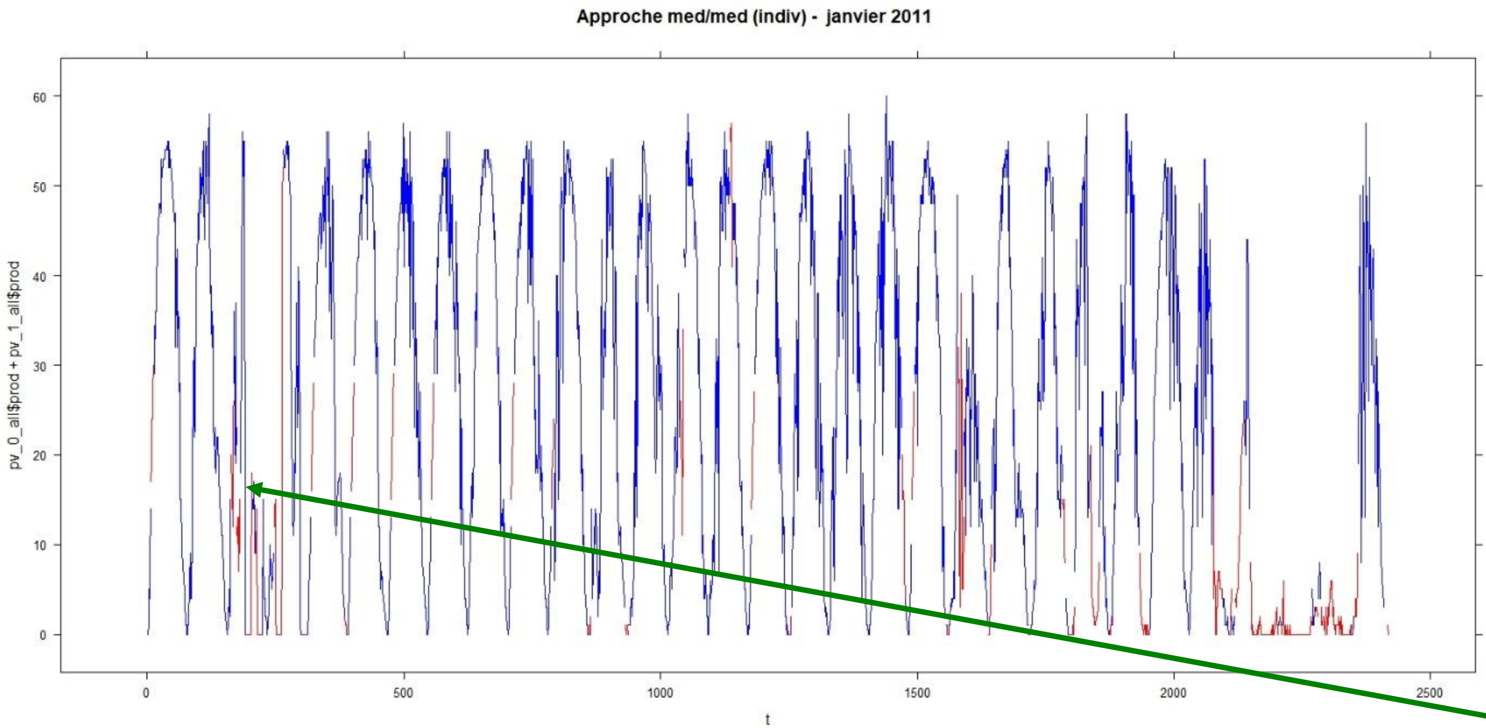
Photovoltaic data (PV stations)



Application of approach on PV data

Photovoltaic data (PV stations)

Are there abnormal hourly PV productions on January 3, 2011 ?

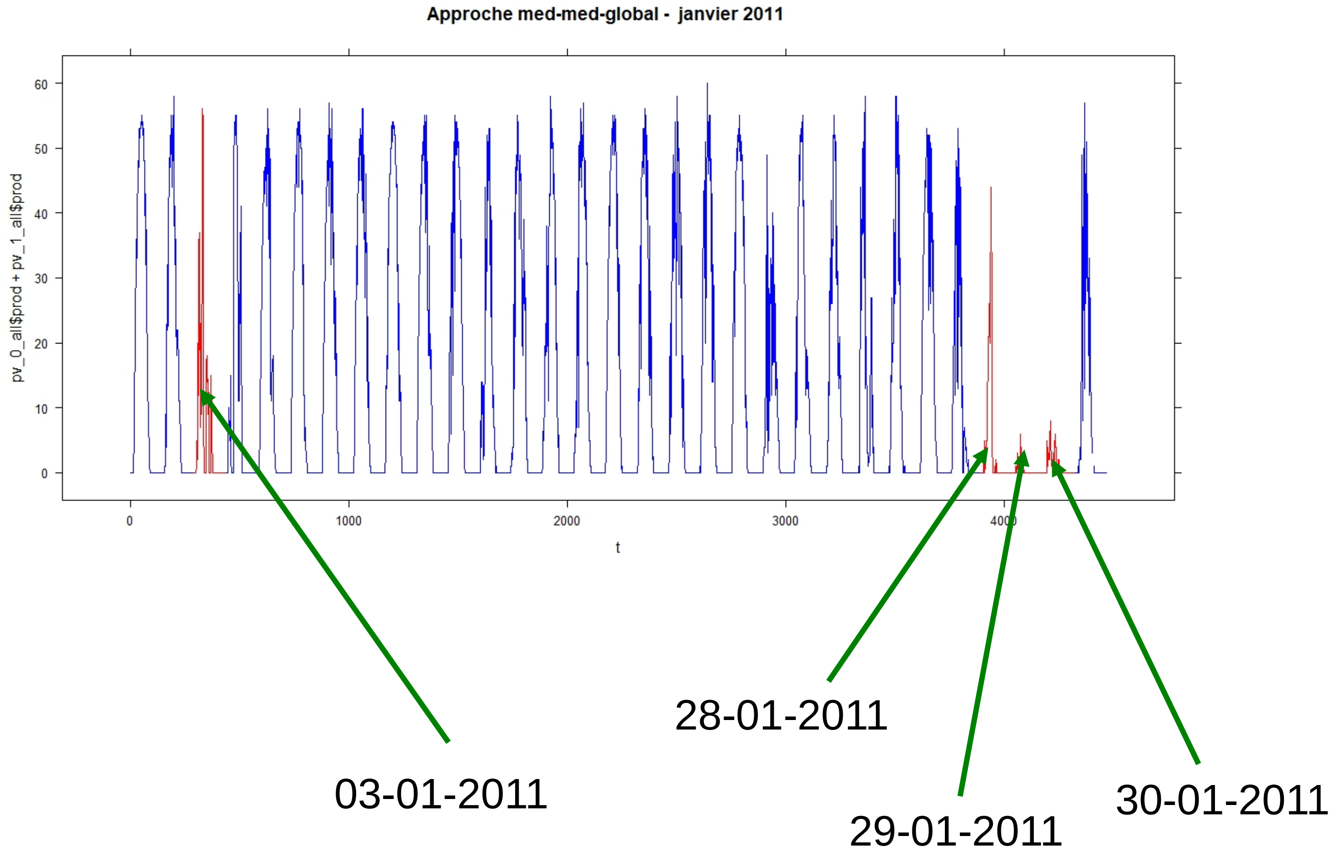


For instance, $s_{j,13}^{(obs)} = 0.0092$ then p-value $= 0.045$ (3-01-2011 – 13:00)

Application of approach on PV data

Photovoltaic data (PV stations)

Overall, is the PV production for that same day abnormal ?

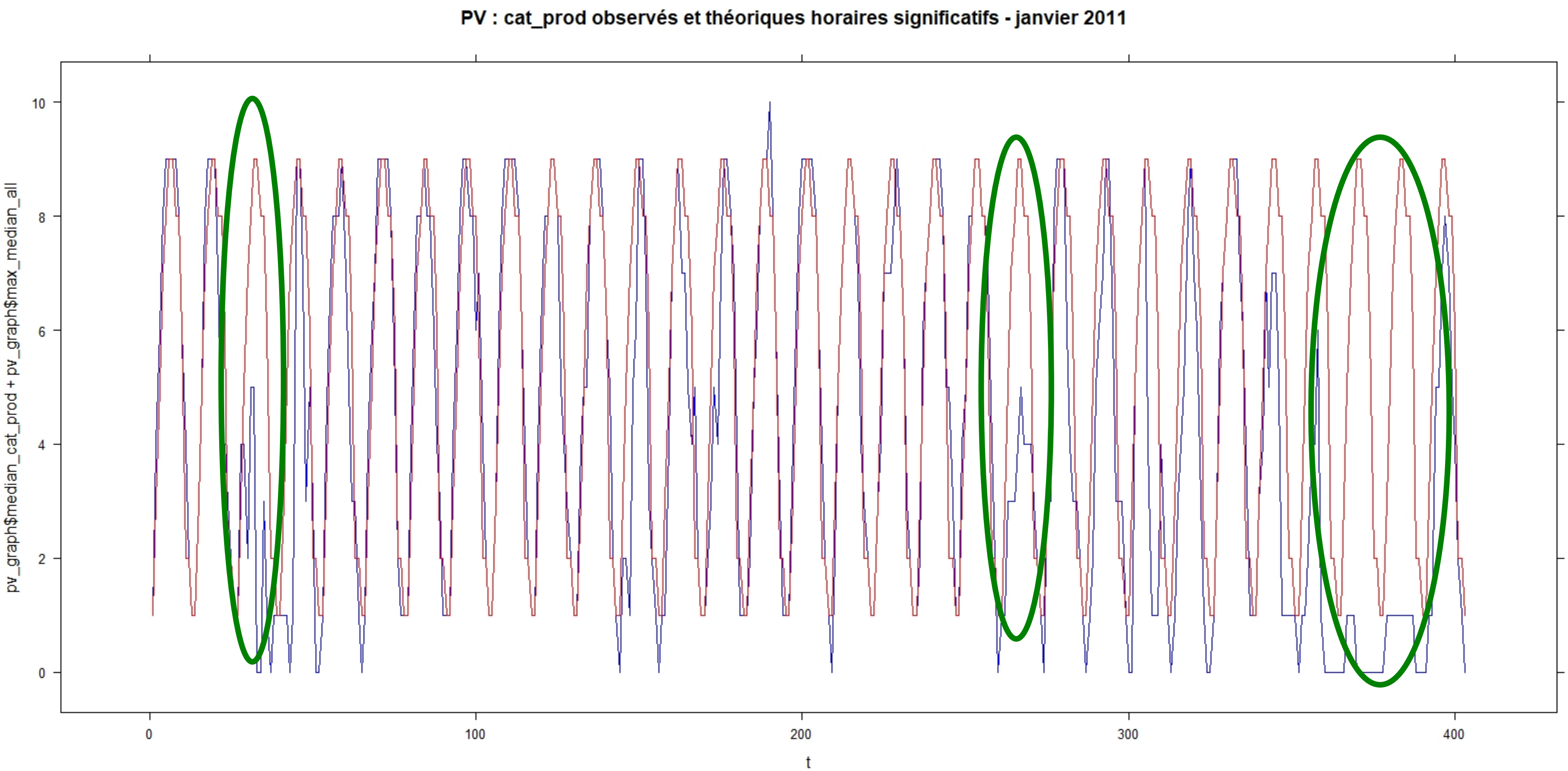


Days	*p*-values
03-01-2011	0,0391
28-01-2011	0,0018
29-01-2011	0,0001
30-01-2011	0,0018

Application of approach on PV data

le cnam

Photovoltaic data (PV stations)



Plan

Context and goal

A non parametric approach for anomalies detection

Application of approach on PV data

$\boxed{\text{Contributions, applications and further researches}}$

Contributions, applications, further research

Approach for regular (or almost regular) time series

Applications : forcasting of PV production, temperature (response or predictor), electric consumption (response), detecting cardiac desease on ECG, …

Comparing to other approaches of anomalies detection

To extend this approach for online strategies

Bibliography

Basseville M., Nikivorov I. (1995): Detection of Abrupt Changes: Theory and Application, Prentice-Hall, Inc.

Choudhary D., Kejariwal A., Orsini F. (2017): On the Runtime-Ecacy Trade-o of Anomaly Detection, arXiv:1710.04735v1, MZ Inc.

Derquenne Ch., (2015): Detection of similar behaviors and abnormal segments in time series, 16th Conference of ASMDA , Piree, Greece.

Fried R., Gather U. (2007): On rank tests for shift detection in time series, Department of Statistics, University of Dortmund, Germany.

Harle F., (2006): Detection de ruptures multiples dans des series temporelles multivariees : application a l'inference de reseaux de dependance. These de doctorat, Universite Grenoble-Alpes, France.

Maya S., Ueno K., Nishikawa1 T. (2019): dLSTM: a new approach for anomaly detection using deep learning with delayed prediction, International Journal of Data Science and Analytics, 8, 137-164. 6