Forecasting and nowcasting with text as data, DSDM

Ben Seimon, Session 2, 4th Apr 2025

Session overview

- Session 1: Forecasting conflict risk
- **Session 2: Practical introduction - forecasting for panel data**

Learning objectives

1. Understand how to generate train-test splits with panel data
2. Understand how to generate target variables
3. Understand the interaction between (1) and (2)

- Session 3: Generate features and predictions!

About me

- DSDM graduate, July 2022
- Practical experience of forecasting with text data:
  - National conflict risk forecasts
  - ADM2 conflict risk forecasts
  - Forced displacement forecast (Google Trends)
- Other interests:
  - Integration of forecasts into decision-making
  - Reinforcement learning, geospatial data, networks

Things you should and shouldn’t do (IMHO)

- **Do:** Read documentation.
    - Going directly to documentation and source code will improve your skills and save you time.
- **Do:** Start small and build unit-level functions.
- **Do:** Inspect outputs by hand. Painful, but necessary.
- **Don’t:** Forget about exploratory data analysis.
- **Don’t:** Generate every imaginable feature. Have a “structural” mindset when it comes to feature engineering (DePrado, 2020).
- **Don’t:** Think that these sessions only apply to conflict forecasting. The principles you will learn apply across domains (healthcare, finance).

Key principles

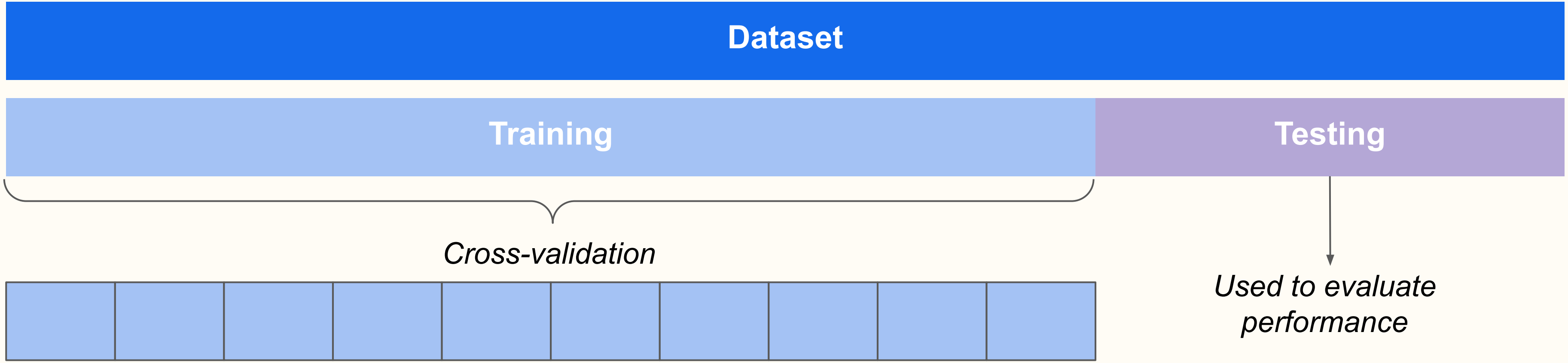
1. Use all available information at time t to train a model that predicts an outcome at:
    a. Nowcasting: T
    b. **Forecasting: $T + h, h > 0$**
2. Update period:
    a. Defines the last period for which you have full information (T)
3. (Train, test) split. A tuple where:
    a. Train = list of indices in the train set
    b. Test = list of indices in the test set

LDA discussion

Forecasting procedure

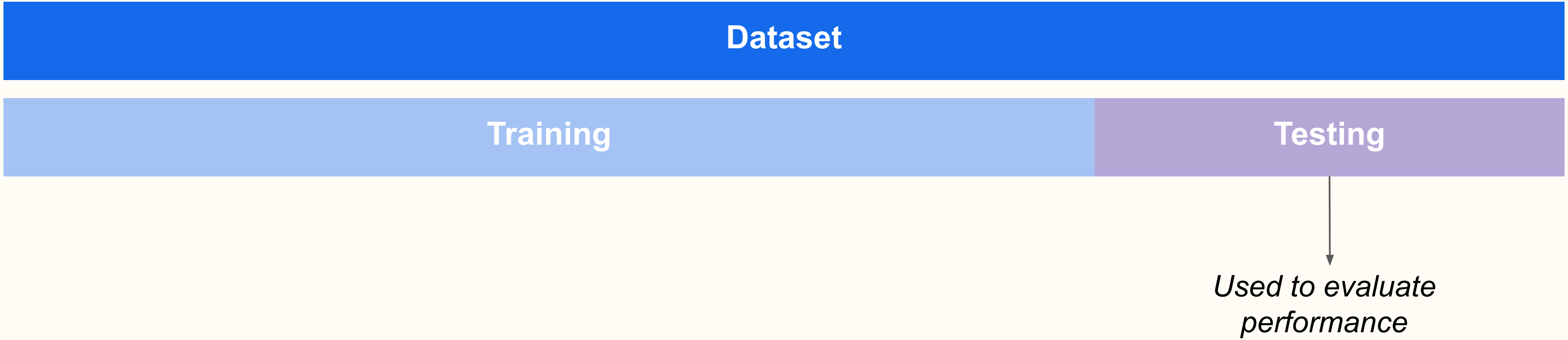
The basics

- ML methods use historical data to learn about the relationship between the features (X) and the outcome (Y)
- Optimally, we want any model that we build to predict outcomes well on future, unseen data
- To mimic this task during model development we use a training and testing set



The basics

- Today I want you to completely forget about cross-validation → it is only relevant for hyperparameter tuning
- Instead the main objective is how to construct a list of valid (train, test) pairs with panel data.



The principle

- Use all available information at time t to train a model that predicts an outcome at:
    - Nowcasting: T
    - **Forecasting: $T + h, h > 0$**

The procedure

Algorithm 1 Rolling Forecast: Pseudo Out of Sample Forecasting

Require: Full data sample $D = \{d_{1989m1}, d_{1989m2}, \ldots, d_{2024m12}\}$
Require: Horizon H
Require: Forecasting model F
Ensure: Forecasts $\hat{Y} = \{\hat{y}_{2010m1<t\leq 2010m1+H}, \hat{y}_{2010m2<t\leq 2010m2+H}, \ldots, \hat{y}_{2024m12<t\leq 2024m12+H}\}$

1: Train model F on data $\{d_{1989m1}, d_{1989m2}, \ldots, d_{2009m12}\}$
2: **for** T from 2010m1 to 2024m12 **do**
3: $D_{\text{train}} \leftarrow$ Data for all $1989m1 \leq t \leq T - H$
4: Retrain model F on D_{train}
5: $\hat{y}_t \leftarrow$ Aggregate forecast of F for $T < t \leq T + H$
6: Append $\hat{y}_t$ to $\hat{Y}$
7: **end for**
return $\hat{Y}$

Train-test splits

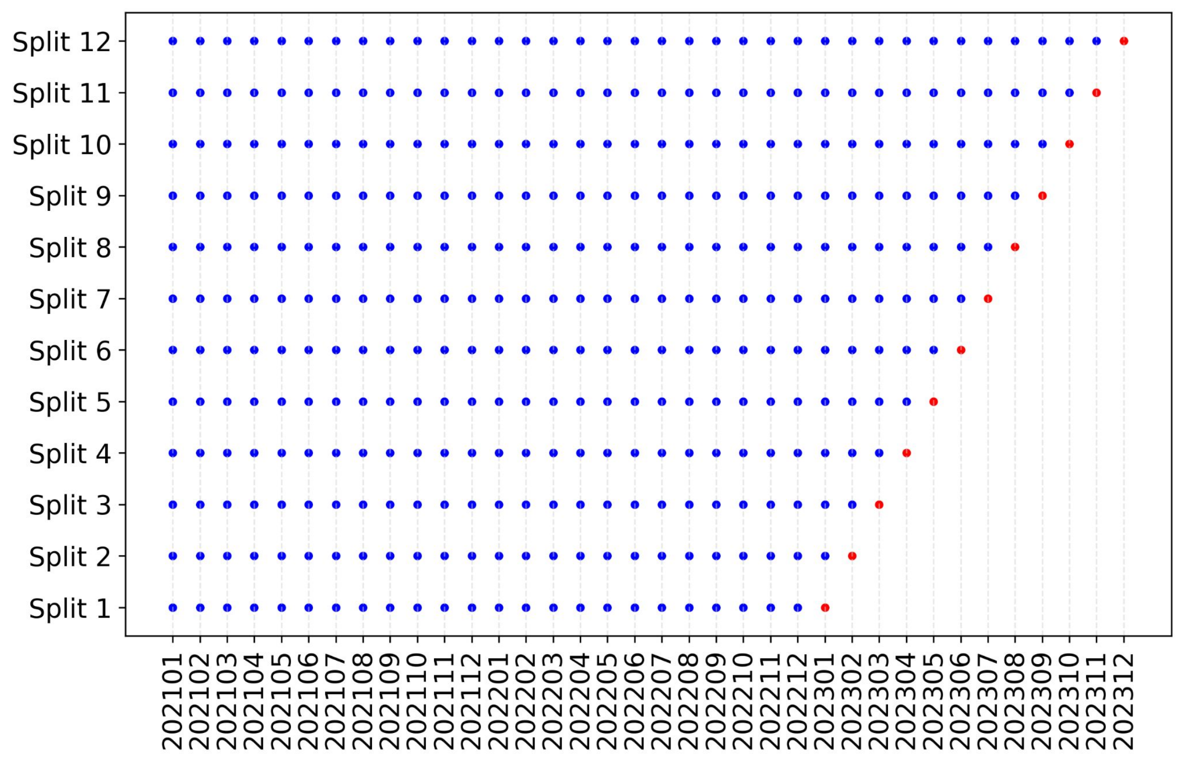
Cross-sectional

- 200 countries (C), 1 timestamp (T)
- Only one split of the data is necessary. A valid (train, test) split would be any of:
  - $([c_1, \dots, c_{150}],\ [c_{151}, \dots, c_{200}])$
  - $([c_{151}, \dots, c_{200}],\ [c_1, \dots, c_{50}])$
  - $([c_{101}, \dots, c_{150}],\ [c_1, \dots, c_{100}, c_{151}, \dots, c_{200}])$
- There is no "forecasting" element here since there is no time dimension.

Time-series

- 1 country (C), 36 timestamps (T)
- Now we must be more careful about how we split the data:
- For simplicity, assume we are only forecasting the next month (*h*=1):
    - $([t_1, \ldots, t_{24}], [t_{25}])$
    - $([t_1, \ldots, t_{25}], [t_{26}])$
    - …
    - $([t_1, \ldots, t_{34}], [t_{35}])$
    - $([t_1, \ldots, t_{35}], [t_{36}])$

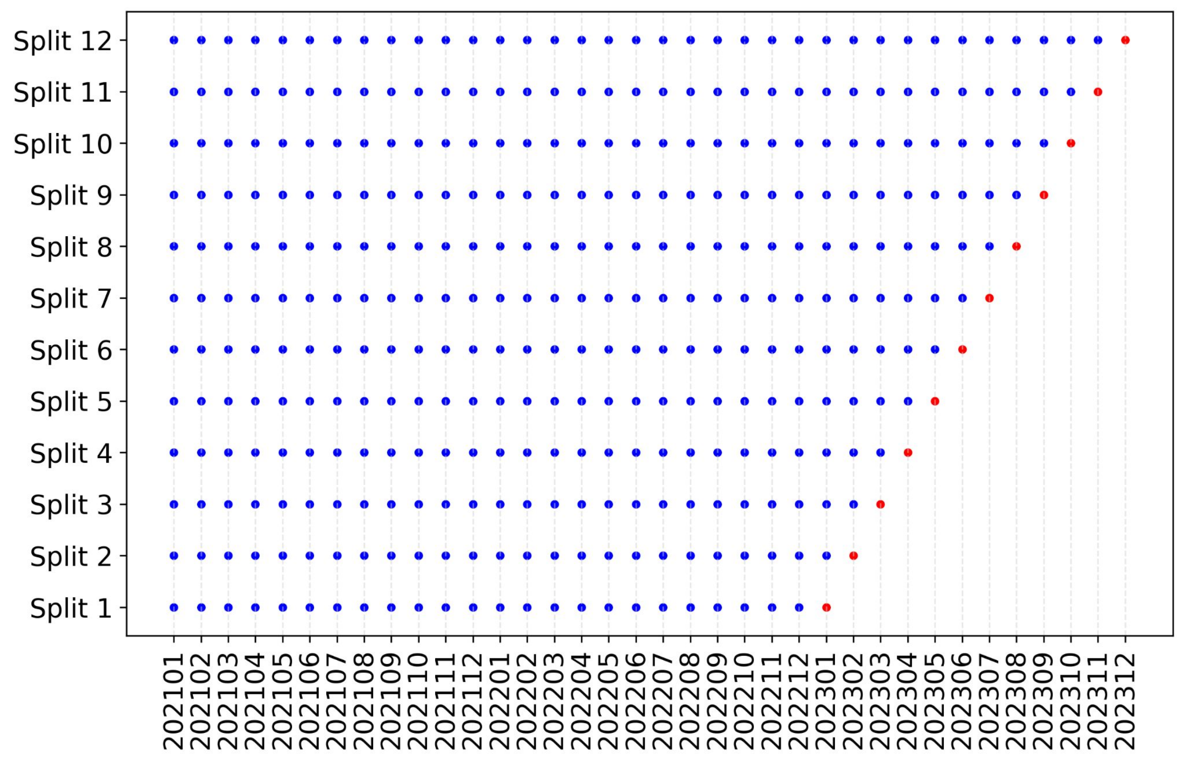
Time-series


Split 12
Split 11
Split 10
Split 9
Split 8
Split 7
Split 6
Split 5
Split 4
Split 3
Split 2
Split 1
202101
202102
202103
202104
202105
202106
202107
202108
202109
202110
202111
202112
202201
202202
202203
202204
202205
202206
202207
202208
202209
202210
202211
202212
202301
202302
202303
202304
202305
202306
202307
202308
202309
202310
202311
202312

Panel data

- 200 countries (C), 36 timestamps (T)
- This is essentially a simple extension of the time-series case:
  - $([(c_1, t_1), (c_2, t_1), ..., (c_1, t_2), (c_2, t_2), \dots, (c_{200}, t_{24})], [(c_1, t_{25}), (c_2, t_{25}), ..., (c_{200}, t_{25})])$
  - $([(c_1, t_1), (c_2, t_1), ..., (c_1, t_2), (c_2, t_2), \dots, (c_{200}, t_{25})], [(c_1, t_{26}), (c_2, t_{26}), ..., (c_{200}, t_{26})])$
  - …
  - $([(c_1, t_1), (c_2, t_1), ..., (c_1, t_2), (c_2, t_2), \dots, (c_{200}, t_{34})], [(c_1, t_{35}), (c_2, t_{35}), ..., (c_{200}, t_{35})])$
  - $([(c_1, t_1), (c_2, t_1), ..., (c_1, t_2), (c_2, t_2), \dots, (c_{200}, t_{35})], [(c_1, t_{36}), (c_2, t_{36}), ..., (c_{200}, t_{36})])$

Panel data



- This is exactly the same as the time-series case.
- Except each of the blue dots will contain the full set of 200 countries; not just one.

Panel data

- The advantage of using panel data is to enable “learning” across units.
- Imagine you treat all 200 countries as individual time-series:
  - Country 50 has no history of violence.
  - You cannot train your model: there are no 1’s in the training data…
- Same applies to a healthcare setting. Imagine you are trying to predict strokes for 20-30 year olds.
  - You want to learn from the rare instances where a person experienced a stroke, and apply that to predict situations for patients with no history.

Target variable

Incidence vs onset

- Our end goal is a target dataframe with:
  - **Index:** [unit, timestep]
  - **Outcome:** The observed outcome for the given unit/timestep
  - **Any_variable:** Binarized version of “outcome”
  - **Inc_variable:** Target variable for incidence
  - **Ons_variable:** Target variable for onset

The “any” variable

- Let’s assume a binary variable called “anyviolence”

isocode	period	violence	anyviolence_th0
AFG	201001	344	1
	201002	536	1
	201003	407	1
	201004	503	1
	201005	502	1
...	...	...	...
ZWE	202310	0	0
	202311	1	1
	202312	0	0
	202401	0	0
	202402	0	0

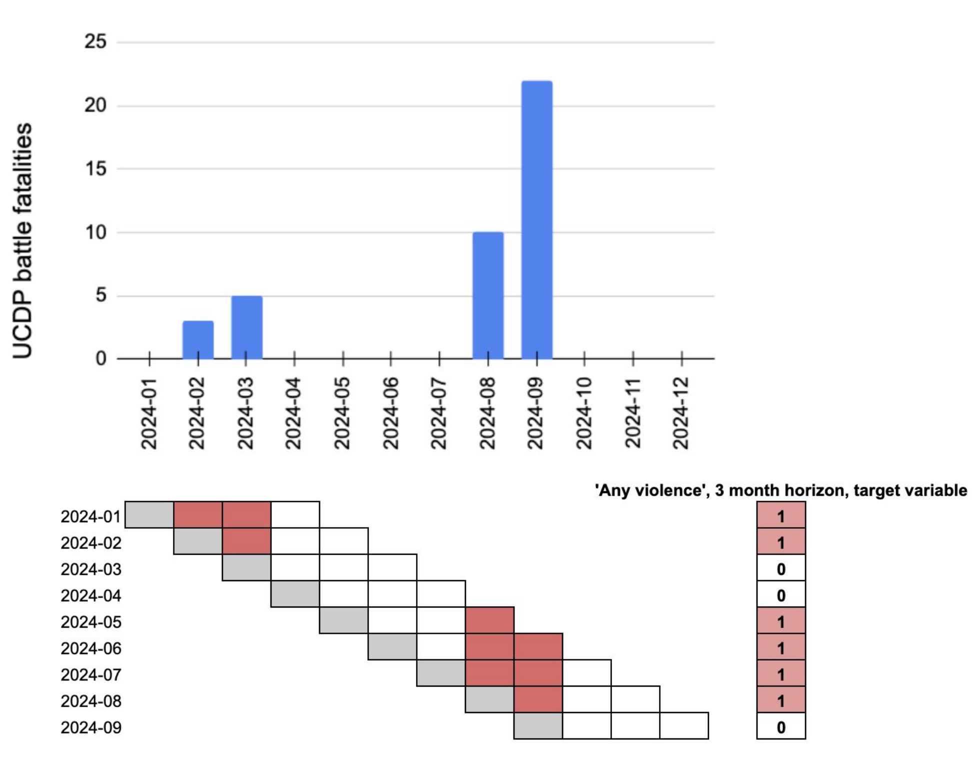
$$\text{any\_violence}(y, \text{threshold}) = \begin{cases} 1, & \text{if } y > \text{threshold} \\ 0, & \text{otherwise} \end{cases}$$

Incidence

- **Incidence:** 1 if anyviolence in the next *h* timesteps; 0 otherwise. Let *h* = 3:

		violence	anyviolence_th0	inc_anyviolence_th0_h3
isocode	period			
LBY	202209	3	1	0.0
	202210	0	0	0.0
	202211	0	0	0.0
	202212	0	0	0.0
	202301	0	0	0.0
	202302	0	0	1.0
	202303	0	0	1.0
	202304	0	0	1.0
	202305	2	1	1.0
	202306	0	0	1.0
	202307	0	0	1.0
	202308	1	1	0.0
	202309	0	0	0.0
	202310	0	0	0.0
	202311	0	0	0.0
	202312	0	0	NaN
	202401	0	0	NaN
	202402	0	0	NaN

Incidence: a visual interpretation



Onset

- **Onset:**
  - 1 if anyviolence in the next *h* timesteps AND no violence in *t*
  - NA if anyviolence in *t*
  - 0 otherwise

Incidence vs onset

isocode	period	violence	anyviolence_th0	inc_anyviolence_th0_h3	ons_anyviolence_th0_h3
LBY	202101	0	0	1.0	1.0
	202102	7	1	1.0	NaN
	202103	2	1	1.0	NaN
	202104	0	0	1.0	1.0
	202105	0	0	1.0	1.0
	202106	3	1	0.0	NaN
	202107	0	0	0.0	0.0
	202108	0	0	0.0	0.0
	202109	0	0	0.0	0.0
	202110	0	0	1.0	1.0
	202111	0	0	1.0	1.0
	202112	0	0	1.0	1.0
	202201	10	1	0.0	NaN
	202202	0	0	0.0	0.0
	202203	0	0	0.0	0.0
	202204	0	0	0.0	0.0
	202205	0	0	0.0	0.0
	202206	0	0	1.0	1.0
	202207	0	0	1.0	1.0
	202208	0	0	1.0	1.0
	202209	3	1	0.0	NaN
	202210	0	0	0.0	0.0
	202211	0	0	0.0	0.0
	202212	0	0	0.0	0.0
	202301	0	0	0.0	0.0

NAs for all situations where any_violence == 1.

There is no "onset" → you are already experiencing violence

Onset vs hard onset

- **Onset:** 1 if anyviolence in the next *h* timesteps AND no violence in *t*; 0 otherwise
- **Hard onset:** 1 if anyviolence in the next *h* timesteps AND no violence in *t-w* <= *t*; 0 otherwise
    - E.g. if we set *w=60*, then we only code a 1 where there is violence in the next *h* timesteps AND where there has been no violence for the last 60 timesteps (including current).
    - N.B. in practice we only use this for evaluation, not the target. You will see this in later sessions.

Train-test splits + target variable

Cross-validation ←→ target variable

- Now things get a little confusing...
- You must always set your target variable based on **known** information
  - Otherwise you are leaking data
- You must replicate this information set for every split

Panel split package

- Luckily for you, there's (now) a package that handles this → panelsplit
- Key parameter:
  - Gap

Last split: what’s wrong here?

- Horizon=3, Threshold=0, **Gap=0.** Forecasting with all information up to the end of 202402. Target is whether there will be any violence in any of 202403/202404/202405.

Last split: Train

isocode	period	inc_anyviolence_th0_h3
AFG	201001	1.0
	201002	1.0
	201003	1.0
	201004	1.0
	201005	1.0
...	...	...
ZWE	202309	1.0
	202310	1.0
	202311	0.0
	202312	NaN
	202401	NaN

Last split: Test

isocode	period	inc_anyviolence_th0_h3
AFG	202402	NaN
AGO	202402	NaN
ALB	202402	NaN
ARE	202402	NaN
ARG	202402	NaN
...	...	...
XKX	202402	NaN
YEM	202402	NaN
ZAF	202402	NaN
ZMB	202402	NaN
ZWE	202402	NaN

Last split: what’s wrong here?

- Horizon=3, Threshold=0, **Gap=0.** Forecasting with all information up to the end of 202402. Target is whether there will be any violence in any of 202403/202404/202405.

Last split: Train

isocode	period	inc_anyviolence_th0_h3
AFG	201001	1.0
	201002	1.0
	201003	1.0
	201004	1.0
	201005	1.0
...	...	...
ZWE	202309	1.0
	202310	1.0
	202311	0.0
	202312	NaN
	202401	NaN

You cannot train on NAs.

Last split: Test

isocode	period	inc_anyviolence_th0_h3
AFG	202402	NaN
AGO	202402	NaN
ALB	202402	NaN
ARE	202402	NaN
ARG	202402	NaN
...	...	...
XKX	202402	NaN
YEM	202402	NaN
ZAF	202402	NaN
ZMB	202402	NaN
ZWE	202402	NaN

First split: what’s wrong here?

- Horizon=3, Threshold=0, **Gap=2.** Forecasting with all information up to the end of 202303. Target is whether there will be any violence in any of 202304/202305/202406.

First split: Train

isocode	period	inc_anyviolence_th0_h3
AFG	201001	1.0
	201002	1.0
	201003	1.0
	201004	1.0
	201005	1.0
...	...	...
ZWE	202210	0.0
	202211	0.0
	202212	0.0
	202301	0.0
	202302	0.0

First split: Test

isocode	period	inc_anyviolence_th0_h3
AFG	202303	1.0
AGO	202303	0.0
ALB	202303	0.0
ARE	202303	0.0
ARG	202303	0.0
...	...	...
XKX	202303	0.0
YEM	202303	1.0
ZAF	202303	1.0
ZMB	202303	0.0
ZWE	202303	0.0

First split: what's wrong here?

- Horizon=3, Threshold=0, **Gap=2.** Forecasting with all information up to the end of 202303. Target is whether there will be any violence in any of 202304/202305/202406.

First split: Train

isocode	period	inc_anyviolence_th0_h3
AFG	201001	1.0
	201002	1.0
	201003	1.0
	201004	1.0
	201005	1.0
...	...	...
ZWE	202210	0.0
	202211	0.0
	202212	0.0
	202301	0.0
	202302	0.0

You are violating the information set. For 202301/202302 the target should be NA since you wouldn't have known what happened in 202304/202305.

First split: Test

isocode	period	inc_anyviolence_th0_h3
AFG	202303	1.0
AGO	202303	0.0
ALB	202303	0.0
ARE	202303	0.0
ARG	202303	0.0
...	...	...
XKX	202303	0.0
YEM	202303	1.0
ZAF	202303	1.0
ZMB	202303	0.0
ZWE	202303	0.0

First split: life is good!

- Horizon=3, Threshold=0, **Gap=2**

Last split: Train

isocode	period	inc_anyviolence_th0_h3
AFG	201001	1.0
	201002	1.0
	201003	1.0
	201004	1.0
	201005	1.0
...	...	...
ZWE	202307	0.0
	202308	1.0
	202309	1.0
	202310	1.0
	202311	0.0

You no longer have NAs.

Last split: Test

isocode	period	inc_anyviolence_th0_h3
AFG	202402	NaN
AGO	202402	NaN
ALB	202402	NaN
ARE	202402	NaN
ARG	202402	NaN
...	...	...
XKX	202402	NaN
YEM	202402	NaN
ZAF	202402	NaN
ZMB	202402	NaN
ZWE	202402	NaN

Panel split package

- Horizon=3, Threshold=0, **Gap=2**

First split: Train

		inc_anyviolence_th0_h3
isocode	**period**	
AFG	**201001**	1.0
	201002	1.0
	201003	1.0
	201004	1.0
	201005	1.0
...	**...**	...
ZWE	**202208**	0.0
	202209	0.0
	202210	0.0
	202211	0.0
	202212	0.0

You are no longer violating the information set.

First split: Test

		inc_anyviolence_th0_h3
isocode	**period**	
AFG	**202303**	1.0
AGO	**202303**	0.0
ALB	**202303**	0.0
ARE	**202303**	0.0
ARG	**202303**	0.0
...	**...**	...
XKX	**202303**	0.0
YEM	**202303**	1.0
ZAF	**202303**	1.0
ZMB	**202303**	0.0
ZWE	**202303**	0.0

Notebook

Next session

Please:

- Go through session_2.ipynb to solidify your understanding. Play around with key parameters just as threshold, horizon, test_size, n_splits, gap to improve your understanding.
- Take a look at https://github.com/4Freye/panelsplit/blob/main/panelsplit/application.py. Here you will see that with a correctly initialized PanelSplit class, predictions are as simple as ps.cross_val_fit_predict()!
- We will go through feature engineering and how to generate predictions!

References

- De Prado, M. M. L. (2020). Machine learning for asset managers. Cambridge University Press.
- https://github.com/4Freye/panelsplit