doi:10.5194/egusphere-egu2020-15729

Evaluating the accuracy of equivalent-source predictions using cross-validation

Leonardo Uieda[1] and Santiago Soler[2]

[1] Department of Earth, Ocean and Ecological Sciences, SOES, University of Liverpool, UK

[2] CONICET, Argentina | Instituto Geofísico Sismológico Volponi, Universidad Nacional de San Juan, Argentina

@leouieda

EGU General Assembly 2020

COMPUTER-ORIENTED GEOSCIENCE LAB

Equivalent source processing

- Powerful tool in potential field data processing
- In a nutshell:
    - Fit a linear model to the data
    - Usually with damped least-squares
    - Linear model = coefficients for a set of point sources
    - Use model to predict data
    - Gridding, upward continuation, derivatives, reduction-to-the-pole, etc.

Controlling parameters

Prediction quality depends mostly on:

1. Damping parameter
2. Depth of sources
3. Location and number of sources

   (see next talk by Santiago Soler | EGU2020-549)

How can we **automatically**
set these parameters
to maximize accuracy?

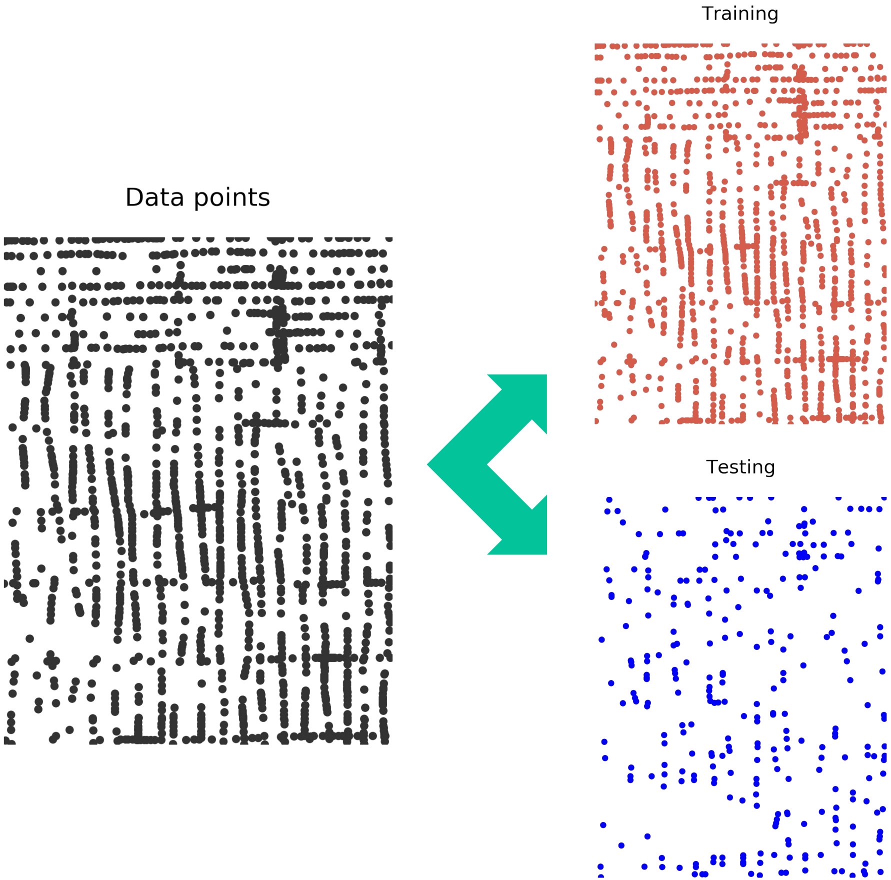
Tuning parameters

Several approaches can be used (some are automated):

- Trial and error (interpreter bias)
- Analytical solutions (difficult to generalize)
- L-curve (limited to damping parameter)
- **Cross-validation** (widely used in machine learning)

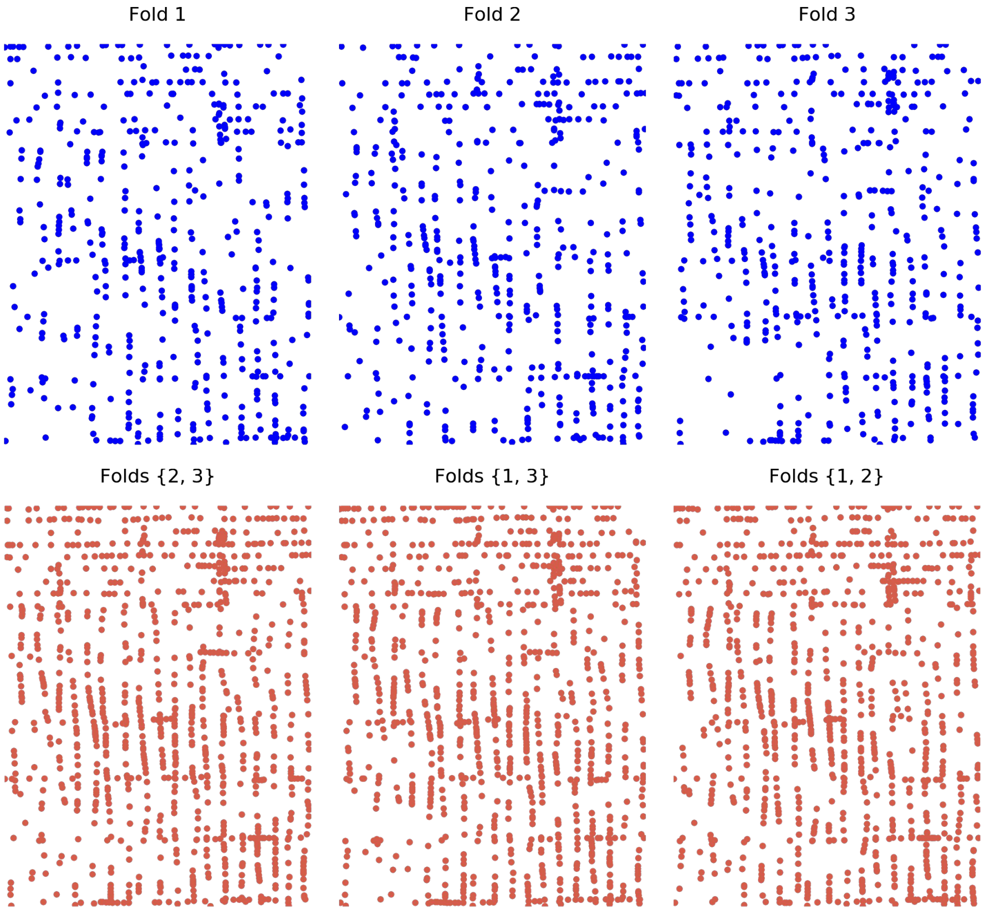
Cross-validation

- Given a set of parameters:
    - Separate data into training and testing
    - Fit on training data
    - Score on testing data (usually mean square error [MSE])
    - Repeat



How to split

- The k-fold method:
  - Separate into *k* parts (folds)
  - Use *i*-th fold for testing
  - Use the rest for training
  - Repeat for *i* in {1, ..., k}
  - Mean of the *k* MSEs

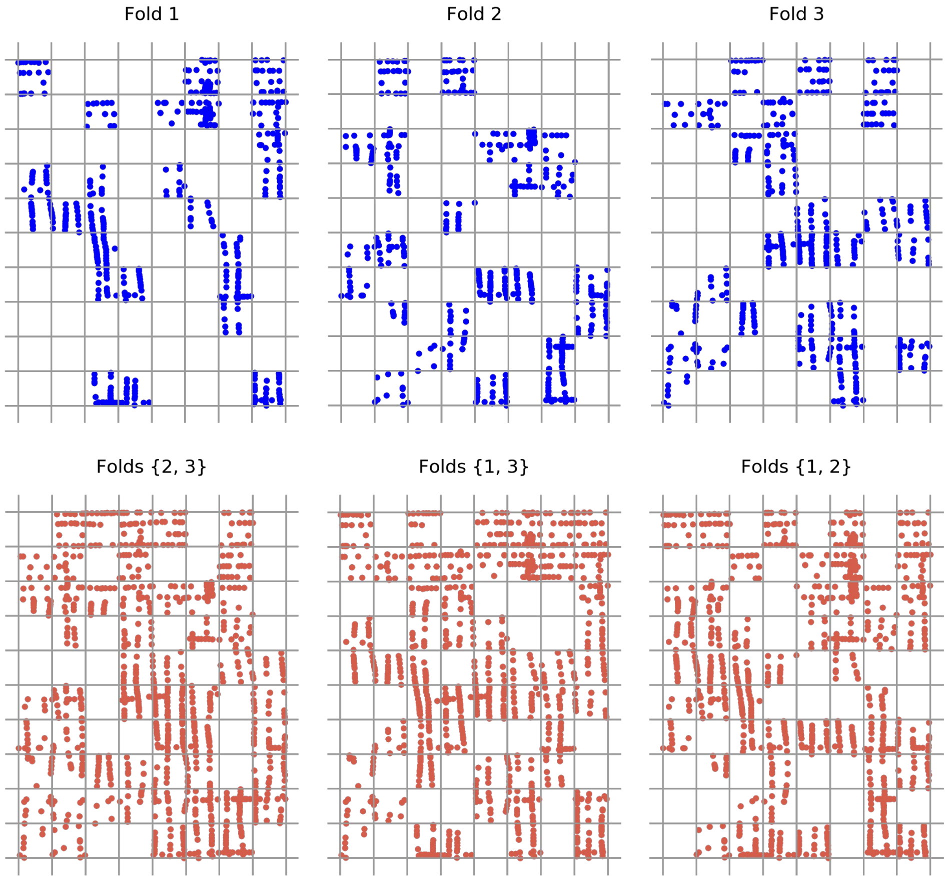


Splitting spatial data

- Measurements tend to be spatially **autocorrelated**
- Evidence from ecology (Roberts et al., 2017) suggests that MSE is **underestimated** by cross-validation
- Solution is to use **blocked** cross-validation (Roberts et al., 2017)
- Guarantee some **distance** between training and testing points

Block k-fold

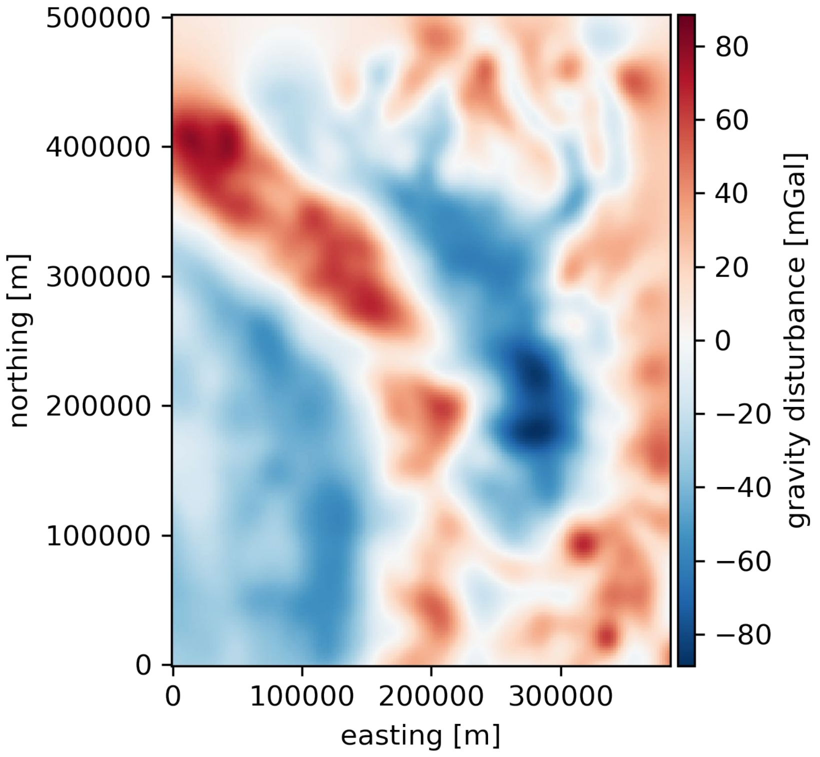
- Split data into blocks
- Split *blocks* into *k* folds
- Assign folds to training and testing like in regular k-fold



Does cross-validation in **equivalent source processing** also underestimate the mean square error (MSE)?

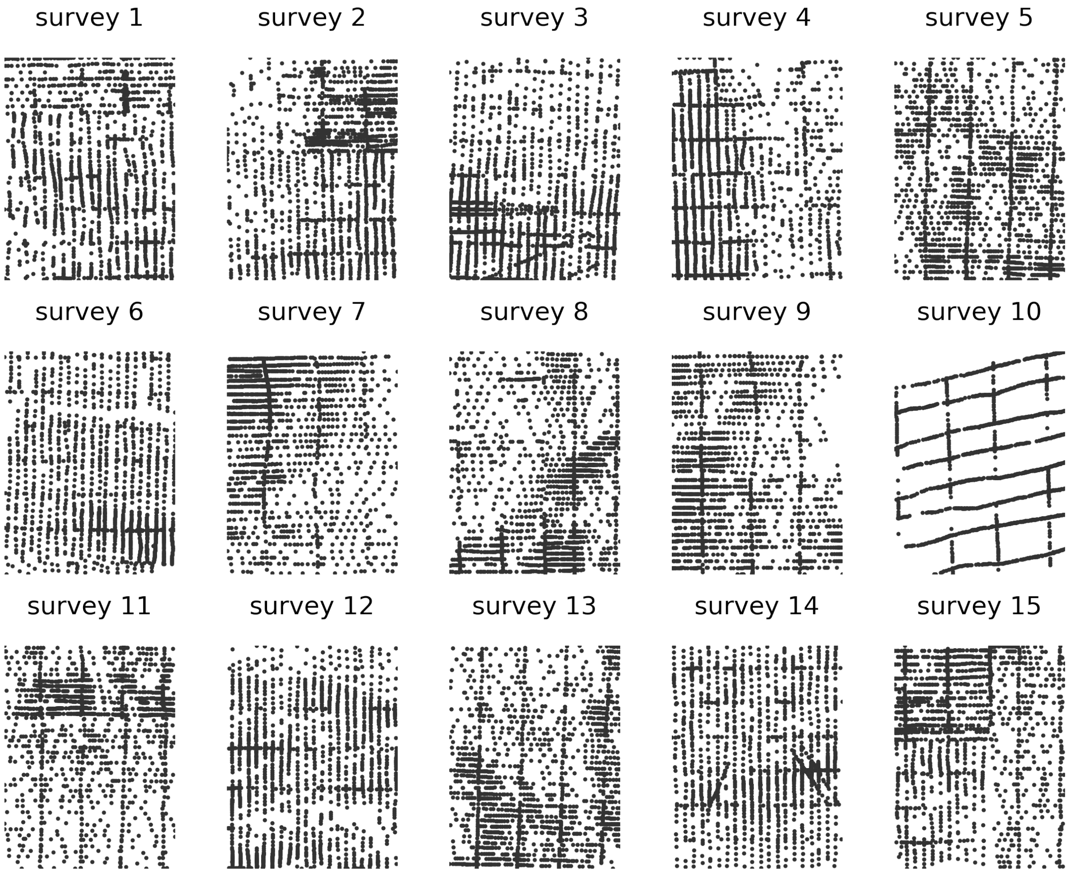
Synthetic data

- We used synthetic data to answer this question
- Simulated based on GRAV-D data (block PN02)
- Grid is **ground truth** for tests



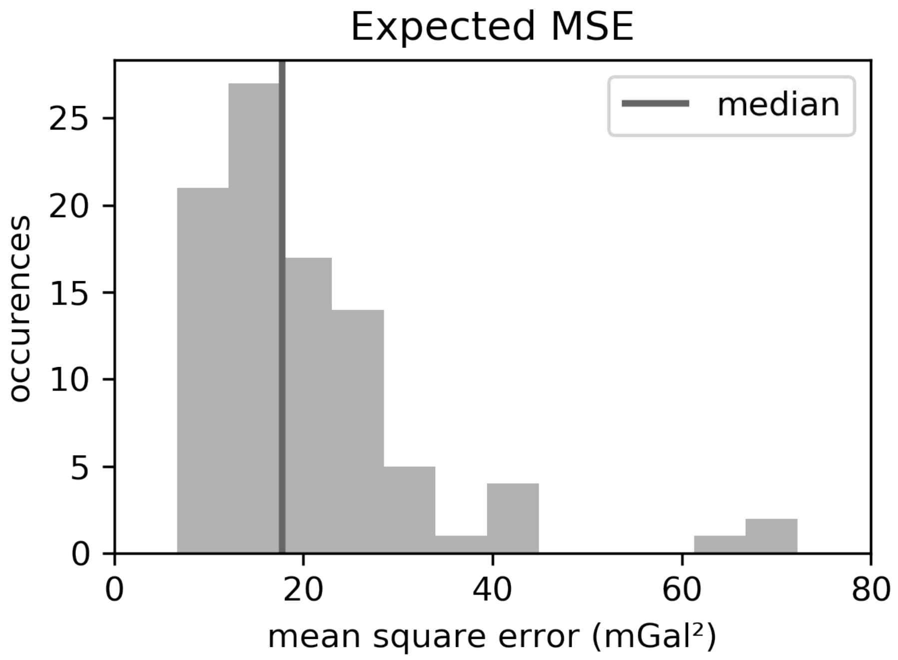
Synthetic surveys

- Sample 100 synthetic surveys from the grid (ground truth)
- Each survey has 1500 data points



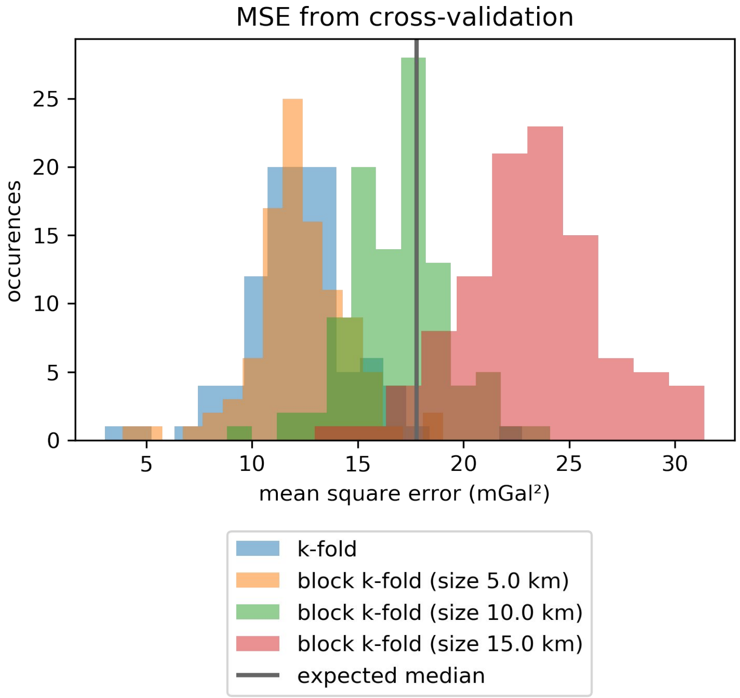
Expected MSE distribution

- Fit equivalent sources on each of the 100 surveys
- With the same parameters (regularization, depth, etc)
- Calculate MSE against **ground truth** (the grid)



MSE estimated by cross-validation

- MSE from cross-validation for the 100 surveys
- k-fold and small block k-fold underestimate expected median
- Large blocks overestimates MSE
- Blocks of 10 km are optimal



Conclusions

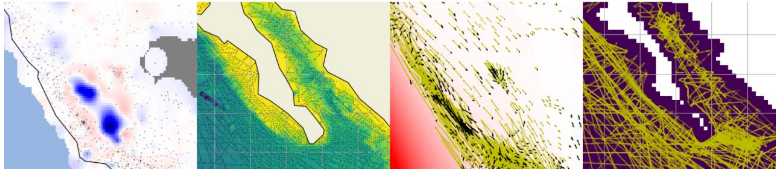
- Cross-validation is an **automated** way to tune parameters
- Scores are **underestimated** when using traditional methods
- **Blocked** cross-validation can solve this issue
- How to choose the **block size**?
- Variogram analysis seems to be the answer (Roberts et al., 2017)
- Paper in progress

Open-source implementation

- Blocked cross-validation will be available in next release of **Verde** (fatiando.org/verde)
- You can try it right now! (install the development version)

VERDE

Processing and gridding spatial data in Python



References

- Roberts et al. (2017): https://doi.org/10.1111/ecog.02881
- GRAV-D Team (2018). "Gravity for the Redefinition of the American Vertical Datum (GRAV-D) Project, Airborne Gravity Data; Block PN02". Available 2020-05-05. Online at: http://www.ngs.noaa.gov/GRAV-D/data_PN02.shtml