Uses and abuses of statistics in geography

David G Rossiter

罗大维教授

南京师范大学地理学学院
(Nanjing Normal University, Geographic Sciences Department)

Section of Soil & Crop Sciences, Cornell University, Ithaca NY USA

ISRIC-World Soil Information, Wageninen, the Netherlands

d.g.rosssiter@cornell.edu; david.rossiter@wur.nl

November 6, 2018

Contents

What do we mean by "statistics"?

- **Descriptive** statistics 描述性统计: numerical **summaries** of **datasets** 数据集
  - Minimum, maximum, range, median, quantiles, histograms, scatterplots . . .
  - "20% of the **samples** 样本 had heavy metal values greater than the legal limit for polluted soils"
- **Inferential** statistics 统计推断: quantitative **statements** about some **population** 全部
  - with **uncertainty** 不确定
  - '20% (±5% one standard error 标准误差) of the **study area** has soils with heavy metal values greater than the legal limit for polluted soils"

Components of an empirical-statistical model

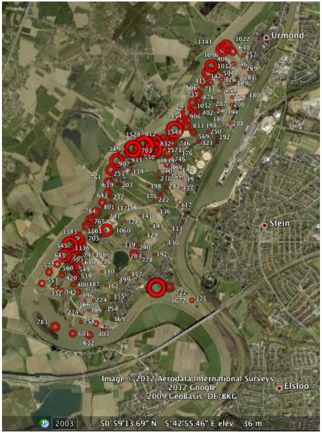
"Empirical-statistical" 经验统计: parameterized equation: dependent ~ independent variables

1. **Predictand**, "dependent" variable 因变量
   - known at **calibration** 校准 observations (locations, times . . .)
2. **Predictors** 独立变量, "independent" variables 自变数
   - for model building 模拟建造, also known at calibration observations
   - for prediction 预测, also known at prediction locations, times . . .
3. **Model form** 模拟类型 relating predictors and predictand
4. **Model parameters** 模拟参数 from calibration 模拟校准
5. **Model evaluation** 模拟评价 : fitness for use

Example dataset – Meuse River (NL) heavy metals 荷兰默兹河重金属数据

predictand log(Zn) 锌对数 concentration in topsoil

predictors: (1) distance to river; (2) elevation

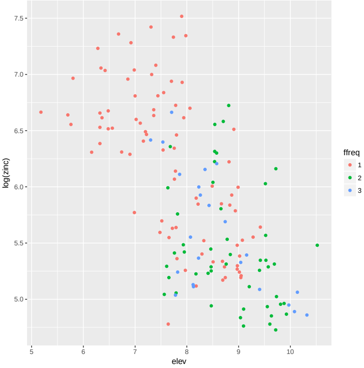


Possible research questions

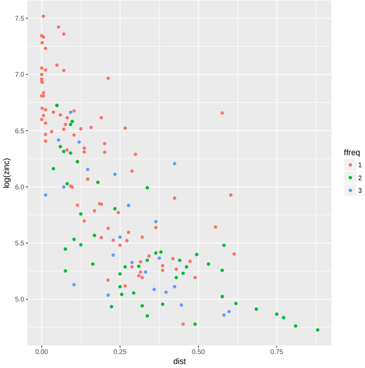
1. **What proportion** of the study area has heavy metal concentrations over regulatory thresholds? 管理限制 → limits land use
2. **Where** are the polluted areas? → **map** = **predict** at unsampled locations
3. **What are the sources** 根源 of the metal?
   - Atmospheric deposition (e.g., from smelters 熔炉)?
   - River floods 洪水?
     - Pre-industrial, from parent rock 母质 upstream
     - Post-industrial, from industry upstream
   - (Soils are from river alluvium 冲积层, none from bedrock 基岩)

Relation of predictand to predictors

elevation (m.a.s.l.)



distance to river (m)



linear?

as square root?

Example empirical-statistical model

Multiple linear regression 多重线性回归, coefficients determined by (“fit by”) Ordinary Least Squares (OLS) 普通的最小二乘法

```
lm(formula = log(zinc) ~ elev + sqrt(dist), data = meuse)

Residuals:
     Min       1Q   Median       3Q      Max
-0.99144 -0.22853  0.00209  0.22244  0.98324

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.64157    0.25206  34.284  < 2e-16 ***
elev        -0.23217    0.03426  -6.777 2.54e-10 ***
sqrt(dist)  -1.97766    0.16025 -12.341  < 2e-16 ***

Multiple R-squared:  0.7226, Adjusted R-squared:  0.7189
```

Interpretation

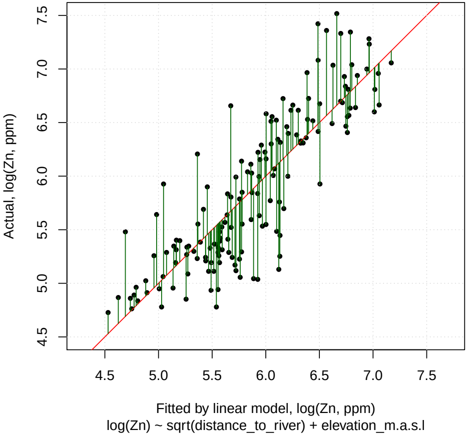
1. An **additive linear model**
   - log(Zn) changes **linearly** with elevation and **linearly as the square root** with distance to river
     - supports the theory that the heavy metal orginates in flood water (higher, further from river → less pollution)
   - no interaction 相互独立 between predictors
2. **Residuals** 残差 are lack of fit; almost ±1 log(Zn)
3. **Coefficients** 系数 show the effect of each predictor; each has a **standard error** (uncertainty)
4. Model **explains** 71.9% of the total varibility 总变量 in the sample set

Is this a correct model?

1. Are the relations between predictors and predictand **linear**?
2. Are the relations **independent** of each other, or are there **interactions** 相互独立 between predictors?
3. Are the assumptions of linear modelling satisfied?
   - Residuals must be independent and identically-distributed residuals; as a group normally-distributed
   - Homoscedasctic 同方差 (same variance across range of predictand)
   - No relation between fitted values 计算值 at observed points and residuals
   - **No spatial or temporal correlation** 相关 among residuals
4. How sucessful is the model for **prediction**?

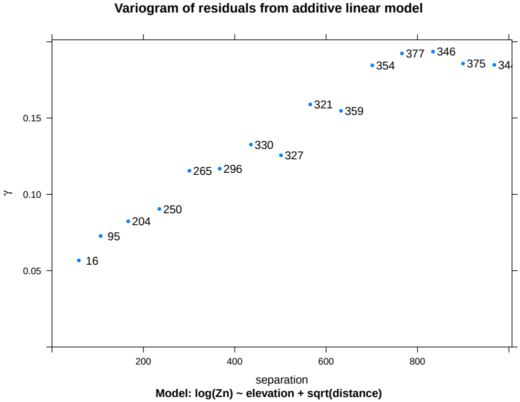
Residuals

Compare on a 1:1 (Actual:Fitted) line.



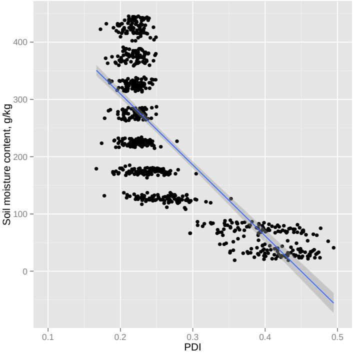
No relation between fitted values and residuals – good!

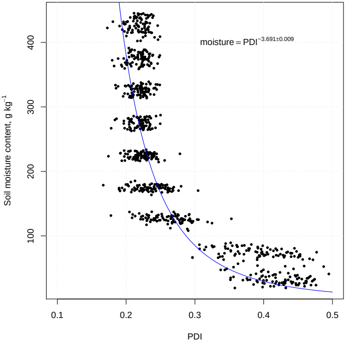
Spatial correlation of linear model residuals



Residuals are *not* independent! Closer separation in *geographic* space → closer separation in *feature* space
This modelling assumption is not satisfied!

Example: Moisture content of surface soils vs. drought indexfrom remote sensing





$R^2 = 0.585$ $R^2 = 0.445$

Linear model is incorrect! although it explains more of the variation. **The relation is obviously not linear**.

Q: Is the power curve 幂函线 shown in the right figure a correct model? (Does it correctly represent the **process** 过程?)

Why do we build statistical models?

1. To (partially) **understand** 理解 (gain insight into 深刻了解) a **geographical process** 地理的过程
   - The **form** of the model suggests the form of the process
   - The **parameters** of the model suggest the influence of predictors
   - The **evaluation** of the model suggest how well the model fits the process
2. To **predict** unobserved locations (mapping 地图绘制) or times (forecasting) or cases (future observations)
   - Apply the model to cases or locations or times, if we know the values of the **predictor** variables
   - Predict with **uncertainty** derived from the model evaluation

Empirical-statistical vs. data mining models

Empirical-statistical 实验性的统计模拟 give an **explicit model** 明确的 which can be examined for **insight** into processes *and* for prediction

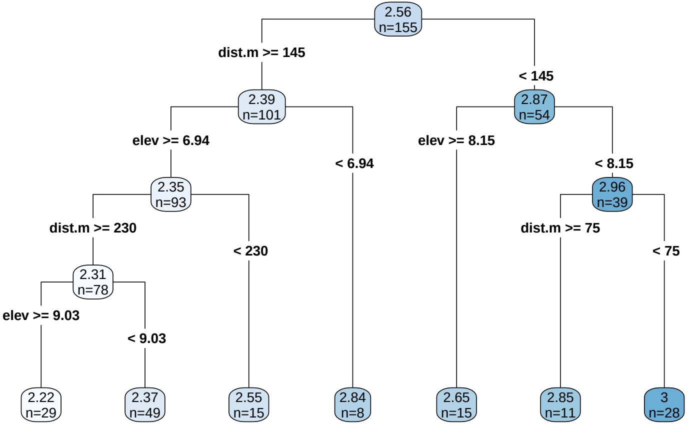
- Examples: multiple (linear) regression, principal components (PCA), discriminant analysis, logistic regression . . .

Data mining 数据挖掘 purely **data driven**, useful for prediction but give little insight into process; “black” (or maybe “grey”) box models

- Examples: Random forests (RF) 随机森林, artificial neural networks (ANN), support vector machines (SVM) . . .

Example data mining model – Regression tree 回归树

Meuse River soil heavy metals dataset

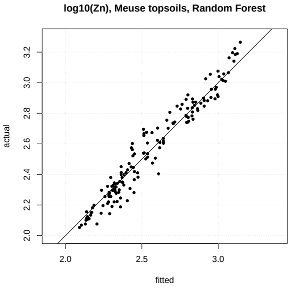


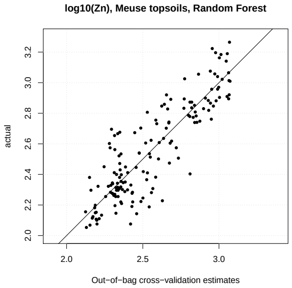
Interpretation

1. 155 observations, mean concentration 平均浓度 log(Zn) = 2.56 mg kg^{-1}
2. Split dataset into two parts, based on distance to river = 145 m; each group with its own mean value
   - $<$ 145: 54 observations, mean 2.87 mg kg^{-1}
   - $>=$ 145: 101 observations, mean 2.39 mg kg^{-1}
   - This is the maximum reduction in within-group variance 组内方差, maximum increase in between-group variance 相组方差
3. Continue to split until improvement in variance is "small"
4. This is purely empirical, putting observations into "boxes", no statistical model is used

Example data mining model – Random forest 随机森林

Use a set of **many trees** with **resampling** 重新取样; predict based on all of these and average them





Calibration fit

Cross-validation fit

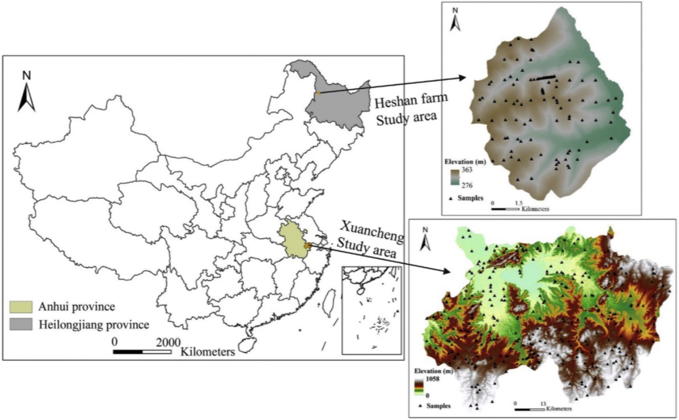
Variable importance 变量重要性 (increase in mean squared error under randomization): flood frequency 9%; distance to river 68%; elevation 54%

Selecting a model form

- Should match what is hypothesized or known about the **process** based on prior knowledge
- Simpler (**“parsimonious”** 吝啬的) is better, don’t complicate a model unless there is a substantial improvement
  - Easier to interpret
  - More likely to give higher **prediction** precision 预测准确
- Compare models based on **evalution statistics**, but don’t change models just on this basis
- Test model form **robustness** 稳健性 by comparing coefficients based on fitting the model with different **random sub-samples**
- **Mapping** models: compare **spatial distribution** 空间分布 of predictions with **landscape** features 地貌特征

Example of competing 比较 model forms

Problem: predict topsoil soil organic matter (SOM) concentration from environmental variables



Source: Zeng, Can-Ying *et al.* (2016). *Mapping soil organic matter concentration at different scales using a mixed geographically weighted regression method*. **Geoderma**:281, 69–82.

Predictor variables

	Variables	Module
Terrain	Elevation	Elevation
	Slope	Slope in ArcInfo
	Planc	Plan curvature (Shary et al., 2002)
	Profic	Profile curvature (Shary et al., 2002)
	TWI	Topographic wetness index (Qin et al., 2009a,b)
	Hand	Height above the nearest drainage (Gharari et al., 2011)
	Dand	Distance to the nearest drainage (Gharari et al., 2011)
	TCI	Terrain characterization index (Park and Van De Giesen, 2004)
	TPI	Topographic position index (Jenness, 2005)
	Flowlen	Flow length based on MFD (Qin et al., 2007)
	ValleyI	Valley index
	RPI	Relative position index (Skidmore, 1990)
	Five binary variables based on fuzzy slope position (Qin et al., 2007) including ridge, shoulder slope (shoulder), back slope (back), foot slope (foot), channel	
Climate	Precipitation	Annual average precipitation
	Temperature	Annual average temperature
Vegetation	EVIs	Summer average EVI
	EVIa	Annual average EVI
Parent materials	Eight binary variables: shale, sandstone, pyroclastic rocks (pyroclastic), granite and granodiorite, limestone, conglomerate, quaternary clay-silt-gravel (clay-silt-gravel), quaternary vermicule boulder and grave clay (grave clay)	

These are all known to affect SOM concentration, via various **processes** → **model** reflects **reality**

Example: higher elevation → cooler temperatures, more rainfall, less evapotranspiration → slower decomposition of SOM

Candidate model forms

1. Multiple linear regression (MLR); select the "best" set of predictors
2. Principal components regression (PCR); predictor set is reduced by Principal Compents Analysis (PCA)
3. Ordinary Kriging (OK): predict only from known points, ignore predictor variables
4. Kriging with an External Drift (KED): MLR with OK of the residuals from MLR
5. Geographically-weighted regression (GWR): like MLR, but coefficients can **vary** across the area
6. GWR-K: GWR with OK of the residuals
7. many more! 等等

Processes implied by these models

1. MLR: covariates **linearly** affect predictand
   - variation: transform predictors or predictands, e.g., log-linear relation
2. PCR: same, but removes colinearity among predictors → identifies latent factors
3. *OK: predictors are not useful, SOM does not depend on the covariates, only local spatial correlation* 空间自相关
4. KED: some variation is explained *globally* as in MLR but residual variation has *local* spatial dependence
5. GWR: covariates linearly affect predictand, but the strength of the relation changes over the area
6. GWR-K: some variation is explained as in GWR-K but residual variation has local spatial dependence

Question: which model best corresponds to the physical process?

- Soil geographers have a well-developed theory:

$$s_0 = f(s, c, o, r, p, a, n)$$

  s_0: soil property to be predicted
  right-hand side: other soil observations climate, organisms, relief, parent material, age, neighbourhood
- But the **functional form** of this equation is not determined by theory
- Many studies of each factor separately (e.g., *chronosequence*, *toposequence* . . .), a few of interactions, none of the complete equation
- We would like our model to correspond to the physical processes by which these factors produce soil.

Comparing mapping results - side-by-side

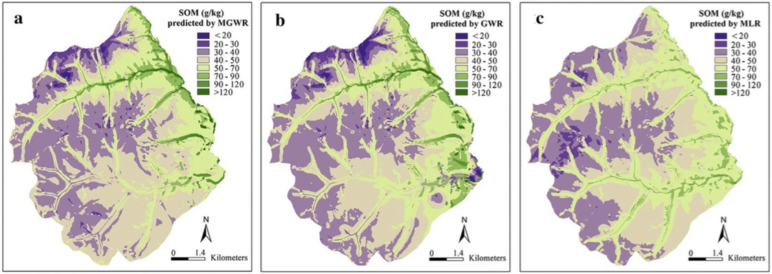
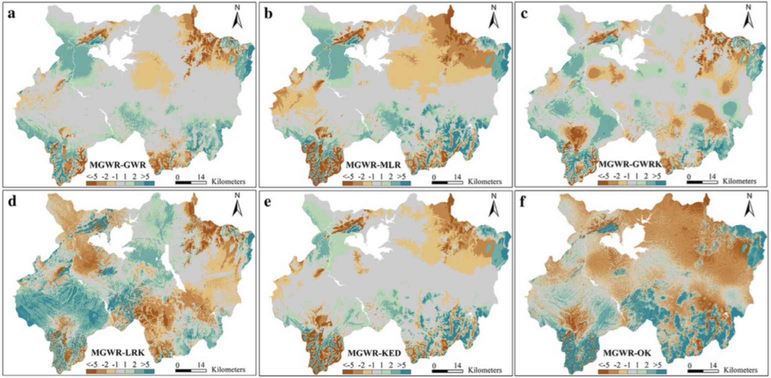


Fig. 7. The distribution of A-horizon soil organic matter maps based on the different model of group 2 for Heshan farm.

Look for **details of soil-landscape** relation.
E.g., small valleys → high moisture → higher SOM.

“Correct” (or at least plausible) spatial pattern → confidence that the model is correct

Comparing mapping results – difference maps



Choice of model form has a large influence on the map!

Evidence that a model form is suitable

1. **internal** 模拟内: from the model itself:
   - how well the model fits the data (success of **calibration**);
   - how well the fitted model meets the model **assumptions**
2. **external** 模拟外 to the model:
   - what is known or suspected about the **process** in the real world that gave rise to the data (what we measure and observe)
     - e.g., atmospheric pressure decreases (linearly ?) with altitude; fewer molecules hold less heat
     - so, we observe cooler temperatures as we move up a mountain
   - how well the model fits observations from the target population
   - success of **evaluation** ("validation" 证实) with:
     - an **independent** dataset
     - a simulated independent set by **resampling** ("cross-validation" 交互证实, "bootstrapping", "jackknifing")

Common abuses and misunderstandings of statistics

1. Not clearly specifying the **population**
2. Making inferences from **unrepresentative samples**
   - Confusing **populations** with **samples**
   - Confusing **descriptive** and **inferential** statistics from a set of observations
3. Internal vs. external model evaluation
   - 1:1 predicted vs. actual, different from linear regression for model evaluation
4. Correlation vs. causation; lurking variables

Abuse: Not clearly specifying the population

- We make **inferences** 推断的结果 about a **population** 全部 from a **sample** 样本 taken from it
  - e.g., map the soil properties in an area, from observations within it
- The sample must cover “all” the variation in the **target** population – the one we want to make statements about.
- **Dangerous to extrapolate** 外推 beyond the limits of the population of which the sample is representative 典型的
  - geographic area; but can argue that the **geographic context** in the extrapolation area matches the calibration area (same climate, same geology . . .)
  - range of measured attributes; no way to know if the relation holds beyond this range

Interpolation 内推 vs. Extrapolation 外推

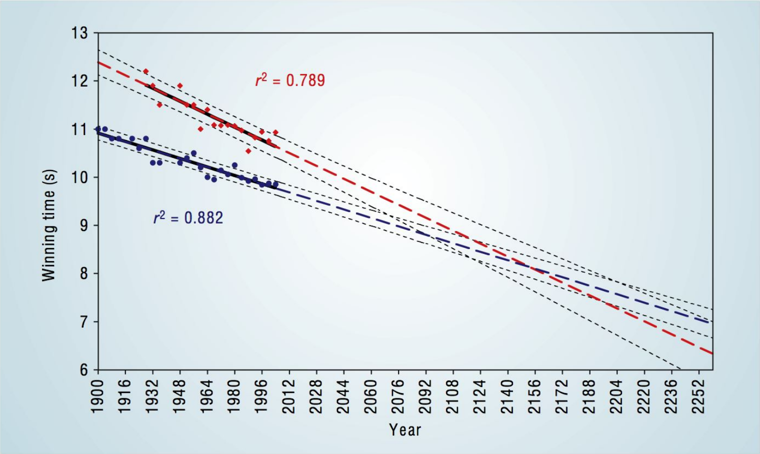
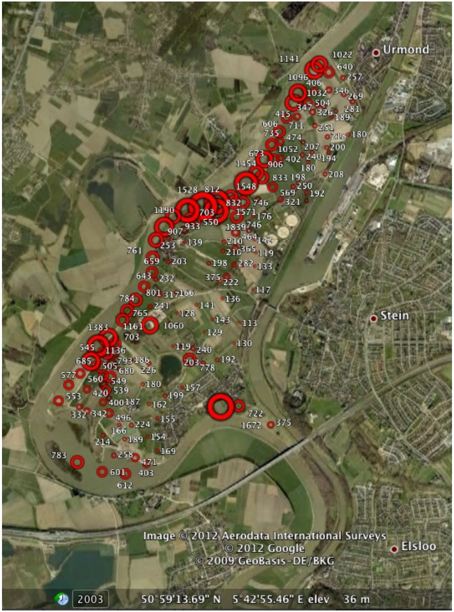


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

Example study area – what is the population?



Purpose:

(1) **map** polluted soils in flood plain of Meuse (Maas) River, NL 荷兰

(2) **infer** source of pollution (upstream industry? aeolian? parent rock?)

Belgium 比利时 on the left bank of river; also flooded.

Is this in the **target population**? Does the sample allow us to say anything about it?

Abuse: Making inferences 推断的结果 from non-probability 不概率 samples

Population a **set** of elements (individuals) about which we want to make a statement

Sample a **subset** of elements taken from a population

What is the relation of the sample to the population?

Relation of the sample to the population

Opportunistic sample 机会样本 See it, grab it; no system; easy access; e.g., sample soils in roadcuts

Purposive sample 故意样本 Select elements based on expert knowledge; e.g.,"typical" ("modal") landscape position to sample soils

"Representative" sample 典型样本 the "expert's" assessment of the purposive sample, no way to check

Probability sample 概率样本 Enumerate all elements that could be sampled ("sampling frame") and use a **random** selection

- completely random, stratified 分层 random, cluster 成群的 sampling, two-stage . . .

Abuse: Confusing the sample and population

Sample "20% **of the samples** had heavy metal values greater than the legal limit for polluted soils"

Population "20% (±5% one standard error 标准误差) of the soils **in the study area** have heavy metal values greater than the legal limit for polluted soils"

The first (sample) is always valid: **descriptive** statistics.

The second (population) is only valid for a **probability sample.** It also allows the computation of **confidence limits** 置信区间 or **credible intervals** 可信区间 for the population.

Population vs. sample statistics

- **Sample**: what was observed – **descriptive** statistics always valid:

```
> summary(meuse$zinc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    113     198     326     470     674    1840
```

- **Population**: what are the true values of the population? Must **infer**.
- Was the sample a **probability sample?** Use a t-distribution of the mean

```
> t.test(meuse$zinc)
95 percent confidence interval: 411.47 527.96
sample estimates: mean of x:  469.72
```

- **Not a probability sample** → use a geostatistical model (other assumptions!)

Abuse: Confusing internal and external model evaluations

Internal 内边 Only using the same **calibration** data that was used to build the model

- Often **too optimistic**; try to minimize by using measures that account for

External 外边 Using independent **evaluation** ("validation") data

- Must be a **probability** sample from the target population
- Can be "pseudo-independent": a simulated independent set by **resampling** (cross-validation, bootstrapping) if the original sample was a probability sample

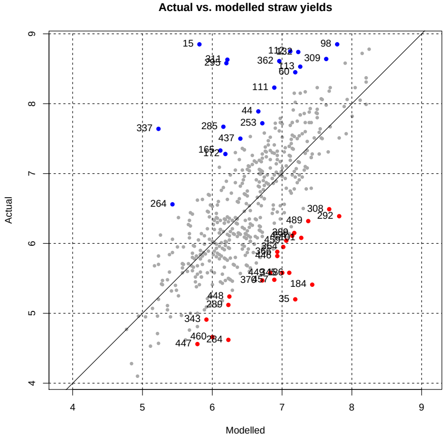
Model calibration vs. evaluation

Calibration 模拟校准 finding “best” values of model parameters

Evaluation 模拟评价 assessing the usefulness of the model for its purpose; fitness for use

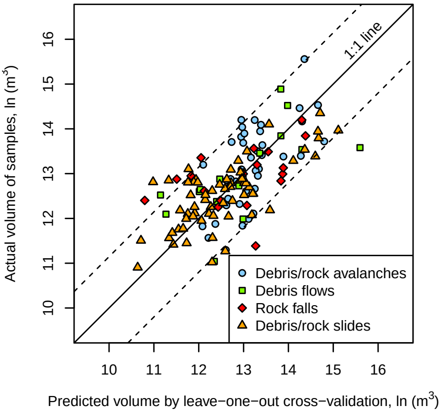
- “Validation” 证实 statistics (RMSE etc.) must be placed in context

Example of internal model fit



Distance to 1:1 line is the **residual** from the fit; i.e., unexplained by model.

Example external evaluation: leave-one-out cross-validation

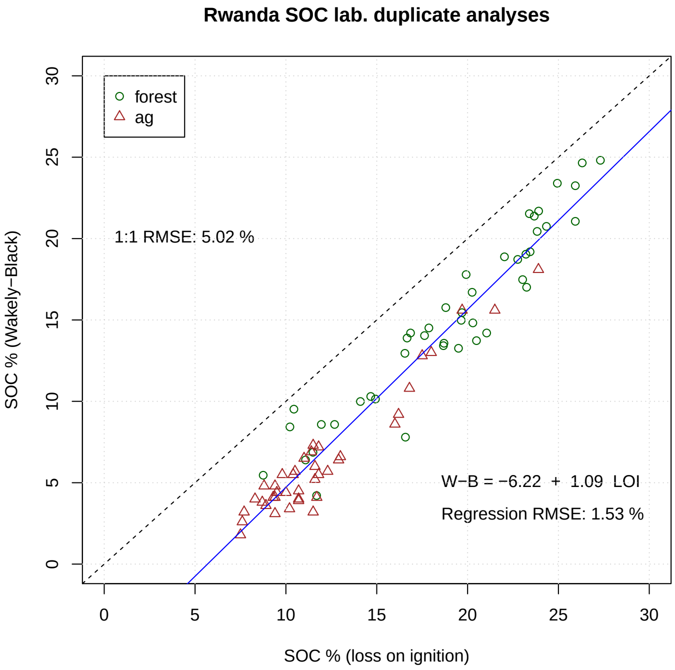


Each observation is predicted by a model built from the other observations.

Abuse: 1:1 vs. linear regression for model evaluation

Compare a model **predictions** with **observations**:

1. **Compare 1:1** 一比一的线 Actual:Predicted 真实的:预测的– this tells how good the model is
   - If cross-validation or independent evaluation sample, a good measure of **predictive precision**
2. **Linear regression** of Actual on Predicted – this tells how much **gain** 增益 and **bias** 偏移 is in the model.


Rwanda SOC lab. duplicate analyses
forest
ag
1:1 RMSE: 5.02 %
W−B = −6.22 + 1.09 LOI
Regression RMSE: 1.53 %
SOC % (Wakely−Black)
SOC % (loss on ignition)
0
5
10
15
20
25
30

Abuse: Correlation/regression 相互关系/回归 vs. causation 原因; lurking variables 潜藏变量

- We may have a good **statistical** relation between one or more predictors X and a predictand (target) y.
- But this does not mean that, in the real world, X causes or influences y.
  - That is a **meta-statistical** argument, from **physical principles** and experiment.
- Example: good correlation between two lab. tests of the same property – does one “cause” the other?
  - No! The “**lurking variable**” 潜藏变量 here is the physical nature of the property itself
  - The statistical relation can be used to translate from one test to the other, with no concept of causation

Example: Correlation vs. causation

- Plant growth modelled as a function of temperature, rainfall and soil nutrient levels – do these **cause** or at least **influence** the plant growth?
  - **Yes**, we know this from lab. experiments.
- Growing season length modelled as a function of crop yield – is there a **causative** relation?
  - **No**, the cause is the other direction. But this model **could be useful** for interpolating growing season in areas with no direct temperature measurements, but where crop yield is measured.

Conclusions

- Statistical models allow us to make **inferences** about **populations**, from **samples** taken from the population
- These inferences include (1) insight into the **processes** in nature; (2) **predictions**
- Models must have an appropriate **form**, be properly **calibrated**, and **evaluated** for their fitness for use

End

(来原:上海气象局)

尽管还难以达到百分之百准确，我们仍要尽百分之百努力。
“Although it is still difficult to achieve complete accuracy, we will still give full effort.”