AESB2440: Geostatistics & Remote Sensing

Lecture 10: Quality of Terrain Analysis Results

Wednesday, May 13, 2015

Roderik Lindenbergh

Contents - Quality

A. Robust statistics

- Median, MAD
- RANSAC

B. Monte Carlo simulation

- Stochastic error propagation
- Slope estimation example

References

Available via Blackboard

A. Robust methods

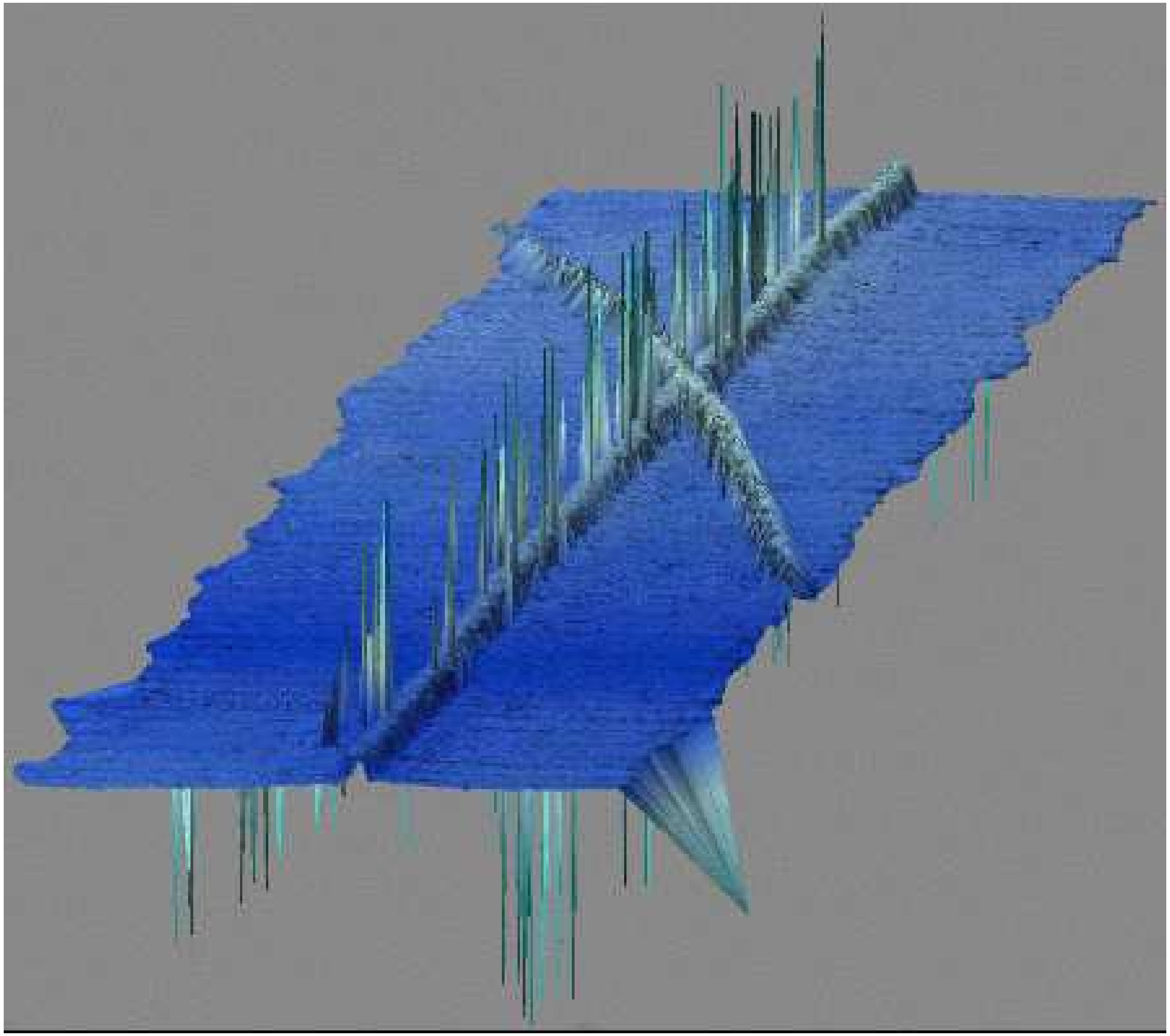
Robust statistics and fitting

Outlier influence

Robust statistics

RANSAC algorithm - line fitting

Outliers

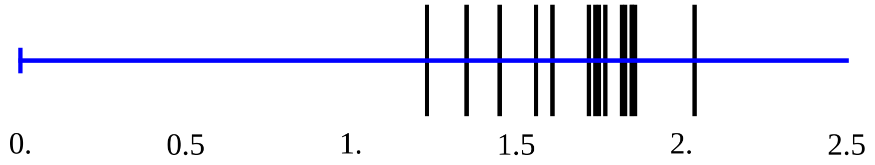


Outliers are points far from the data

Outliers spoil statistics!

Case 1

$$S_1 = \{1.23, 1.35, 1.45, 1.56, 1.61, 1.72, 1.74, 1.75, 1.77, 1.82, 1.83,\ 1.85, 1.85, 1.86, 2.04\}$$

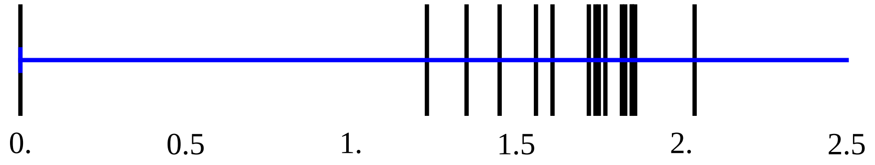


mean S_1 = 1.695

median S_1 = 1.75

Case 2

$$S_2 = \{0, 1.23, 1.35, 1.45, 1.56, 1.61, 1.72, 1.74, 1.75, 1.77, 1.82, 1.83,\ 1.85, 1.85, 1.86, 2.04\}$$



mean S_2 = 1.589

median S_2 = 1.745

Question: why this number of digits in mean and median?

Robust statistical methods

A statistical method is robust if its outcome is not changing dramatically if outliers are added.

Mean, standard deviation: one outlier can spoil the outcome completely

Robust alternatives:

- Median,
- MAD - Median of Absolute Deviations

Median and MAD

Let $S = \{1, 1, 2, 2, 4, 6\}$

$$m_S = \text{median}(S) = 2$$

$$\begin{aligned} \text{MAD}(S) &= \text{median}(\{|S_1 - m_S|, |S_2 - m_S|, \ldots, |S_n - m_S|\} \\ &= \text{median}(\{|1-2|, |1-2|, |2-2|, |2-2|, |4-2|, |6-2|\}) \\ &= \text{median}(\{1, 1, 0, 0, 2, 4\}) \\ &= 1 \end{aligned}$$

Scale MAD to get the equivalent of Standard Deviation.

Quantiles

Let F be a distribution function.

For a random variable X, the *distribution function* is fiven by

$$F : \mathbb{R} \to [0,1],\ \text{s.t. } F(a) = P(X \leq a), \text{for} -\infty < a < \infty$$

Sample Median: $q_n(.5) \approx q_{0.5} = F^{-1}(0.5)$, the distribution median

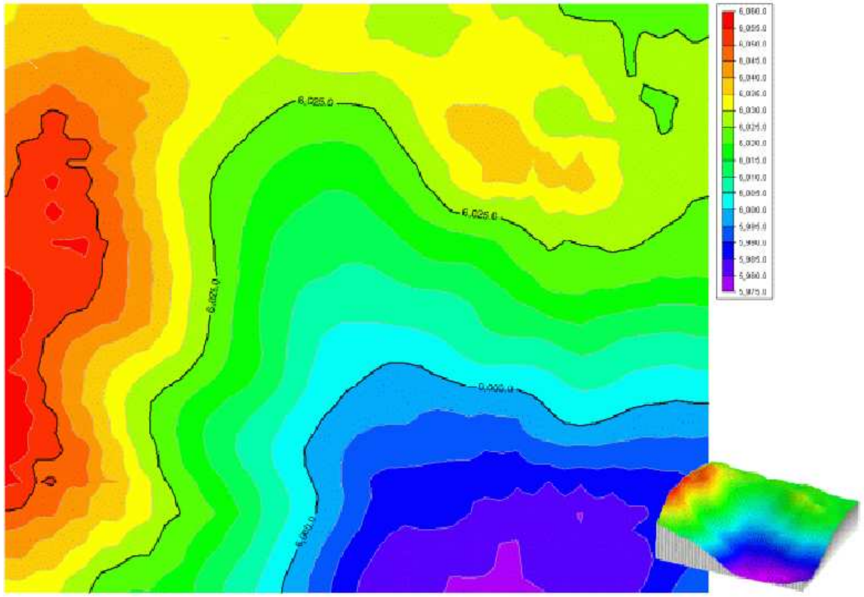
Or, more general,

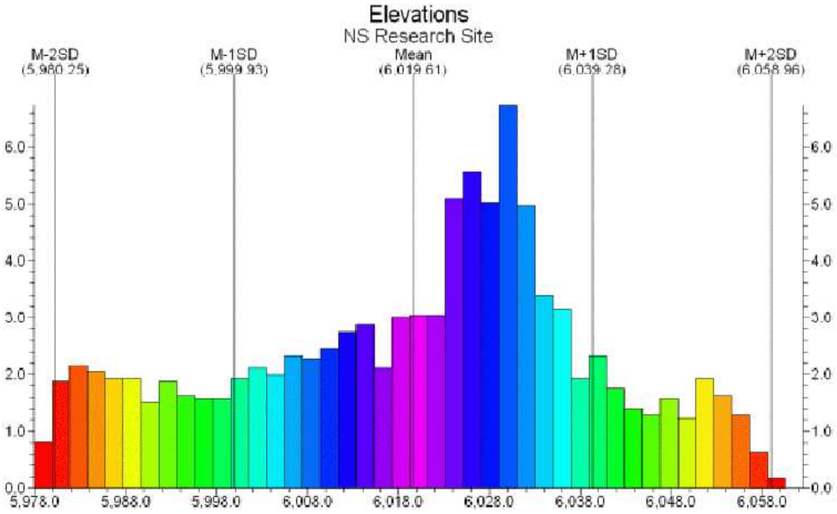
pth empirical quantile: $q_n(p) \approx q_p = F^{-1}(p)$, the distribution quantile

Robust maximum and minimum

Question: How does the expectation relates to the sample median for symmetric distributions?

Question: What are robust alternatives for the maximum and minimum of a (large) data set?





Source https://www.rockware.com/rockworks/revisions/2005$_$q2.htm

MAD vs. Standard Deviation

Claim: Let F be an arbitrary normal distribution $N(\mu, \sigma^2)$ with mean μ and standard deviation σ and let Φ be the standard normal distribution $N(0, 1)$. Then

$$\text{MAD}(F) \ = \ \sigma\Phi^{-1}(.75) \ \approx \ 0.6750\sigma$$

Question. So how to get a robust equivalent of the standard deviaton?

Sketch of the proof of the claim:

1. Let X be a random variable of F, and let m be the median of F. The distribution function of the random variable $Y := |X - m|$ is given by

$$G(y) \ = \ F(m + y) - F(m - y)$$

2. $\text{MAD}(F) = \text{median}(G)$ and

$$G^{-1}(.5) \ = \ F^{-1}(.75) - F^{-1}(.25)$$

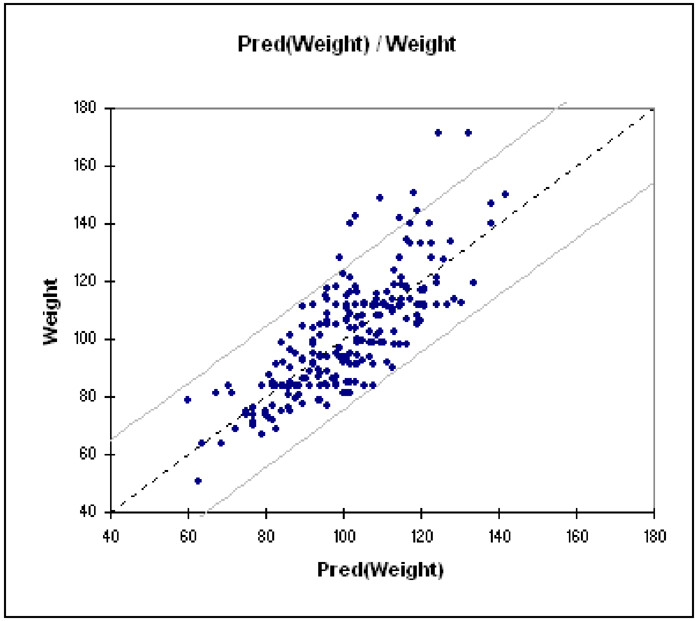
[F.M. Dekker et al.,*A modern introduction to Probability and Statistics*, Springer, 2005]

Outlier removal: Top Down

Mitigating the effect of outliers:

Simplistic Top Down approach:

1. Start with all the data
2. Fit e.g. a line through the data
3. Estimate the st.dev of the line fit
4. Remove all data over 3σ away
5. Go back to Step 1.



Disadvantages??

- ...
- ...

Diagnostic Methods

Top down:
Do your thing and afterwards evaluate where it goes wrong.

More sophisticated method:
Data snooping: stochastic evaluation of outliers

Recall least-squares:

$\underline{y} = \{y_1, \ldots, y_n\}$	Vector of observations
$\underline{\hat{y}}$	Vector of adjusted observation
$\underline{\hat{e}} := \underline{y} - \underline{\hat{y}}$	Vector of Residuals

w-test:
Is the error $\hat{e}_i$ in observation y_i acceptable, given the known quality of this observation?

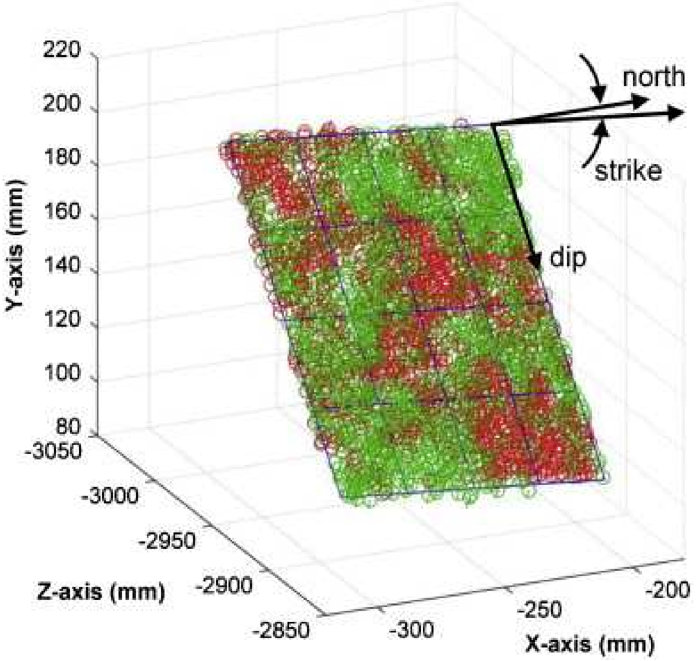
Question: remaining problem with this method?

Bottom up: The RANSAC paradigm

Recall: in many cases there are many more observations then needed to fit a geometric object.

Example
Plane on the right:

- Maybe 100 000 points in 3D
- How many points needed?



RANSAC:

1. Use as little (random) observations as possible for fitting/estimating
2. But repeat that many times
3. Finally select the best option

Ransac - line interpolation

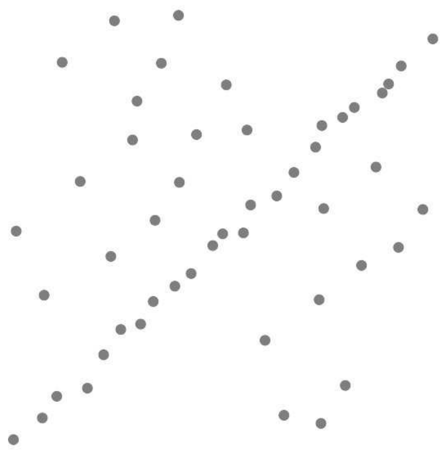
RANSAC - Random Sample Consensus

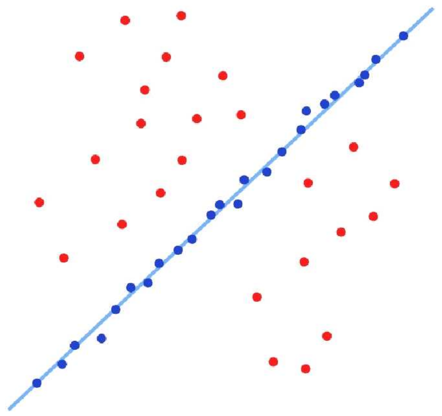
RANSAC for line fitting (example)

1. Select two random points
2. Fit a line through the two points
3. Determine residuals between all points and line
4. Divide points into two classes
   (a) inliers - points with small residual
   (b) outliers - points with large residual
5. Score of the run: number of inliers
6. Return to 1.

Choose, after enough runs, that fit that has the highest score, i.e. the largest amount of inliers

Ransac - line fitting result





Number of iterations ⇔ Number of outliers

Example:
If 50% is inlying, the chance of randomly picking two inliers is 0.25

Note: More parameters needed to parameterize underlying model
⇒ more iterations needed!

RANSAC - general setup

n	number of observations	e.g. laser scanner: $n \approx 100000$
m	number of model parameters	line: $m = 2$
p	probability that observation belongs to model	(nr. of inliers)/n
ϵ	model treshold	maximum distance between observations and model
k	number of trials	should be enough to get a model fit based on inliers only

Number of RANSAC trials needed

$$\begin{aligned} z \quad &:= \quad \text{Prob(at least one trial is outlier free)} \\ &= \quad 1-\ \text{Prob(all trials contain outliers)} \\ &= \quad 1-(1-p^m)^k \end{aligned}$$

$$\Rightarrow \quad (1-z) \; = \; (1-p^m)^k$$

Corollary. In order to ensure one outlier free trial, on average at least k trials are needed, with

$$k \; = \; \frac{\log(1-z)}{\log(1-p^m)}$$

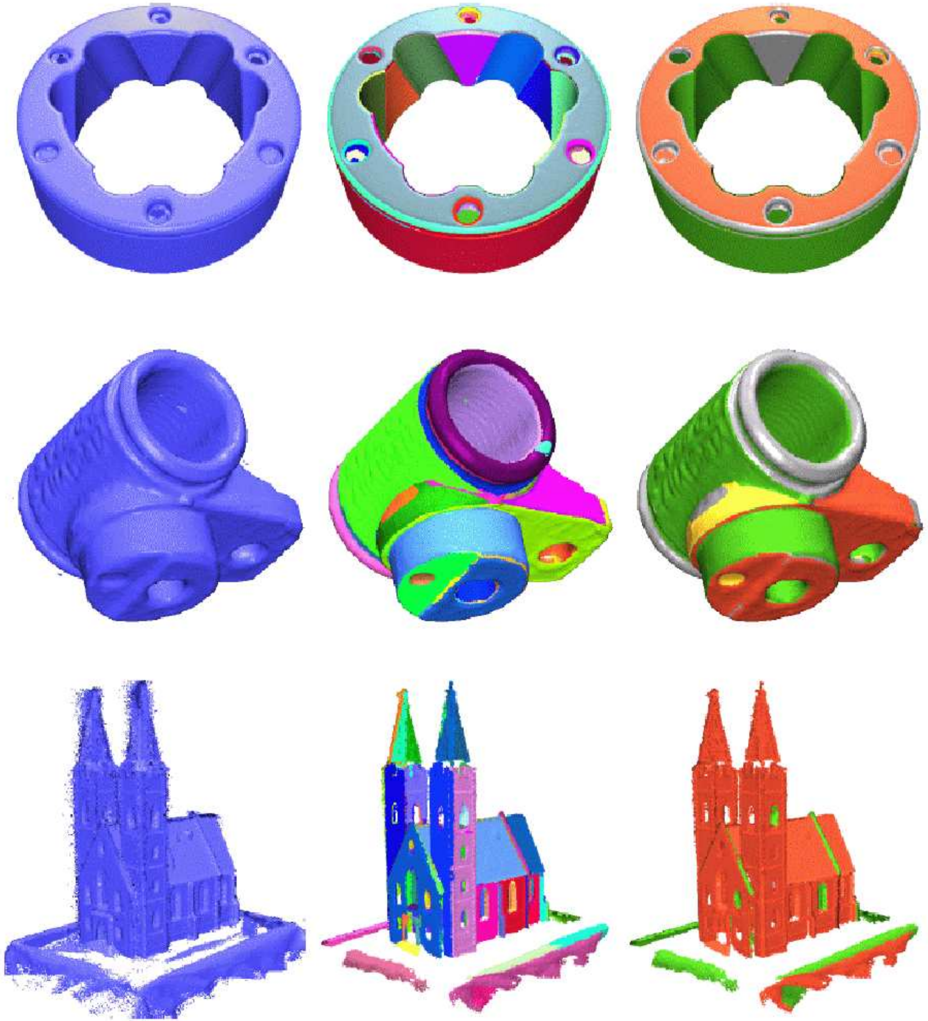
Example. [Line Fitting.] Number of points $n = 50$; Inlier probability $p = 50\%$; Number of model paramaters $m = 2$.

Probability z on an outlier free trial	90 %	99 %	99.9 %
Average required number of runs k	8	16	24

RANSAC for point cloud segmentation

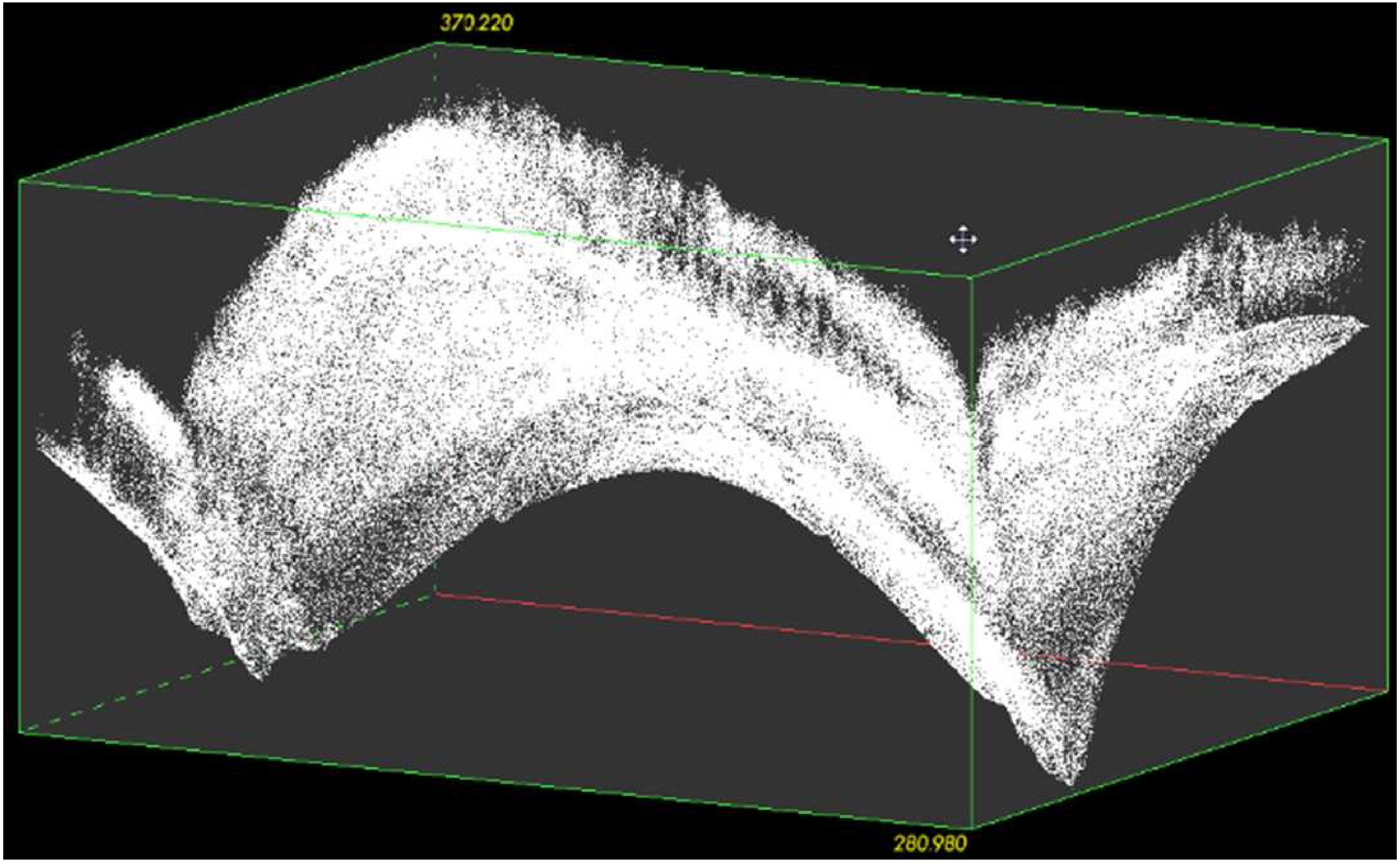
Efficient RANSAC for Point-Cloud Shape Detection,
Ruwen Schnabel, Roland Wahl und Reinhard Klein
In: Computer Graphics Forum (Juni 2007), 26:2(214-226)

http://cg.cs.uni-bonn.de/de/publikationen/paper-details/schnabel-2007-efficient/



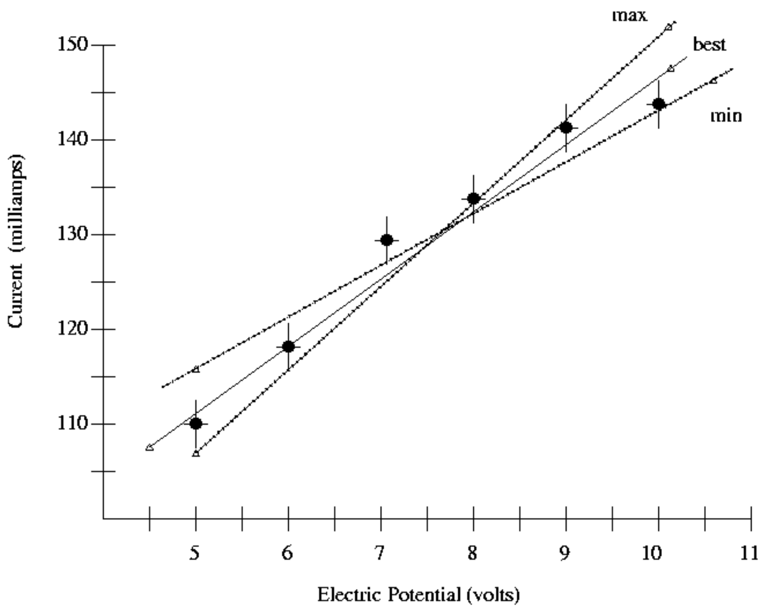
B. Monte Carlo Simulations

Does slope affect tree height estimation?



Source http://lbi-archpro.org/als-filtering/lbi-project/reference-data-set

Propagating uncertainty



Errors in the measurements propagate into errors in estimated parameters.

Example.
Figure: errors in current measurements result in uncertainty in the estimated slope.

http://spiff.rit.edu/classes/phys311/workshops/w2c/slope_uncert.html

Assessing the propagated uncertainty

At least three approaches:

A. Validation against ground truth data.

- Needed: data of better quality
- Often not available, certainly not for derived parameters

B. Formal error propagation

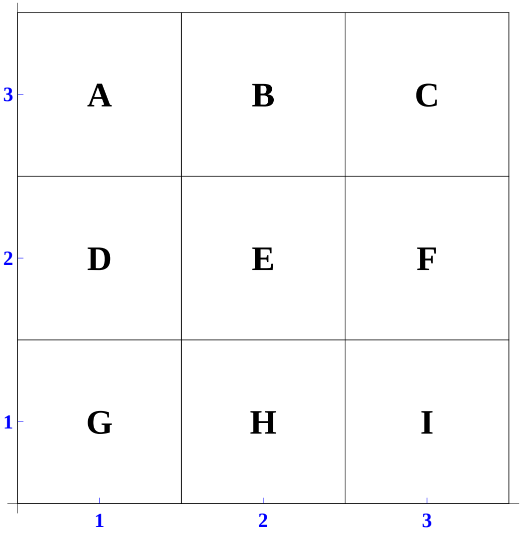
- Example: Kriging variance depends on quality and proximity observations
- Still difficult to get insight in the sensitivity of derived parameters to errors in the observations.

C. Simulating many possible results

- Variation in the outcomes of the results gives insight in the sensitivity
- Possible to directly simulate the needed derived parameters

Example: slope determination

Consider the slope at the middle E of the 3×3 window.



Simple method for estimating slope s:

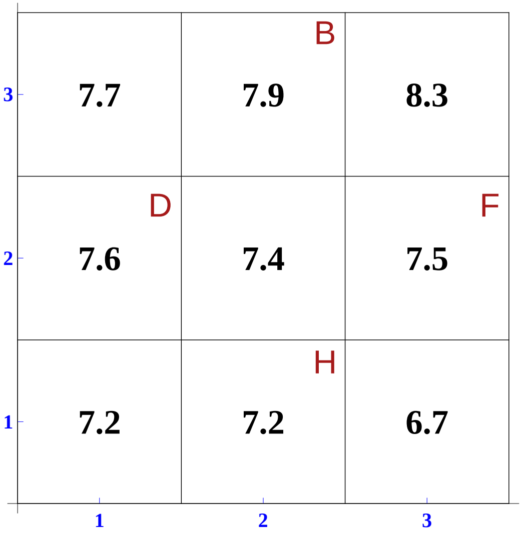
$$\begin{aligned} \nabla_x &\approx \frac{1}{2}(F - D) \\ \nabla_y &\approx \frac{1}{2}(B - H) \\ \nabla &= (\nabla_x, \nabla_y) \end{aligned}$$

$$\begin{aligned} \mathrm{s} &= \|\nabla\| = \sqrt{\nabla_x^2 + \nabla_y^2} \\ &= \frac{1}{2}\sqrt{(F-D)^2 + (B-H)^2} \end{aligned}$$

Question. Other methods for slope estimation?

Example: slope estimation

Suppose we are given the observations on the right.



	1	2	3
3	7.7	7.9	8.3
2	7.6	7.4	7.5
1	7.2	7.2	6.7

Slope estimation based on the observations:

$$\hat{s} = \frac{1}{2}\sqrt{(F-D)^2 + (B-H)^2} = 0.354$$

Question: What is this slope in degrees?

Problem: not clear now how reliable this estimation is.

In situ validation

Possible, but strenuous

Source https://www.youtube.com/watch?v=vlCiJma_rpA

Formal error propagation, example

Assume parameter $g(X)$ depends linearly on random variable X, that is

$$g(X) = rX + s, \qquad \text{with } r, s \in I\!R$$

Assume, moreover, that X is normally distributed with standard deviation σ_X.

Question. What is $E\{g(X)\}$?

Before: $E\{rX + s\} = rE\{X\} + s$

Question. What is the variance of the random variable $Y = rX + s$?

σ_Y^2 = $E[(Y - \bar{Y})^2]$ = $E[((rX + s) - (r\bar{X} + s))^2]\}$ = $E[r^2(X - \bar{X})^2]$ = $r^2 E[(X - \bar{X})]^2$ = $r^2\sigma_X^2$.

Conclusion: for a simple relation $g(.)$ we have propagated the uncertainty in X to the uncertainty of $g(X)$.

Question. Is our relation for the slope ‘simple‘?

Monte Carlo simulation, idea

1) Generate many possible scenario’s of height values,

2) Determine for each scenario the corresponding slope

3) Evaluate the spread of the slope results over the different scenario’s

I'M AFRAID HE CAN'T COME TO THE PHONE, HE'S IN THE MIDDLE OF A MONTE CARLO SIMULATION
Deals Department
BRMM BRMM!
RALLY!
GODDARD

Goddard Cartoon ©PharmaVentures; all rights reserved

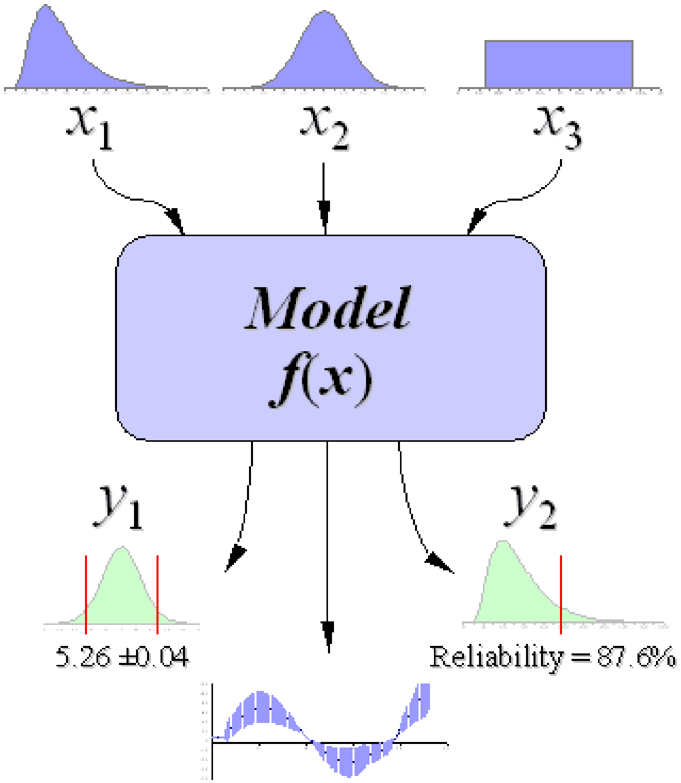
Source http://www.dailytech.com/Detroit+Researcher+Receives+250000+NSF+Grant+for+New+Structural+Failure+Method/articl

Monte Carlo method, input

Input:

A) Relation $v = g(u_1, u_2, \ldots, u_k)$ between observations $u_1, \ldots, u_k$ and derived parameter v.

B) cumulative distribution functions $F_{u_i} : I\!R \rightarrow [0,1]$ for each of the observations $u_1, \ldots, u_k$.



Recall: cumulative distr. function is defined as

$$F(a) = P(X \leq a), \qquad \text{for} -\infty < a < \infty$$

where P stands for Probability, [Dekking et al., 2005]

Source http://www.vertex42.com/ExcelArticles/mc/MonteCarloSimulation.html

Draw from a distribution

Obtaining a distribution:

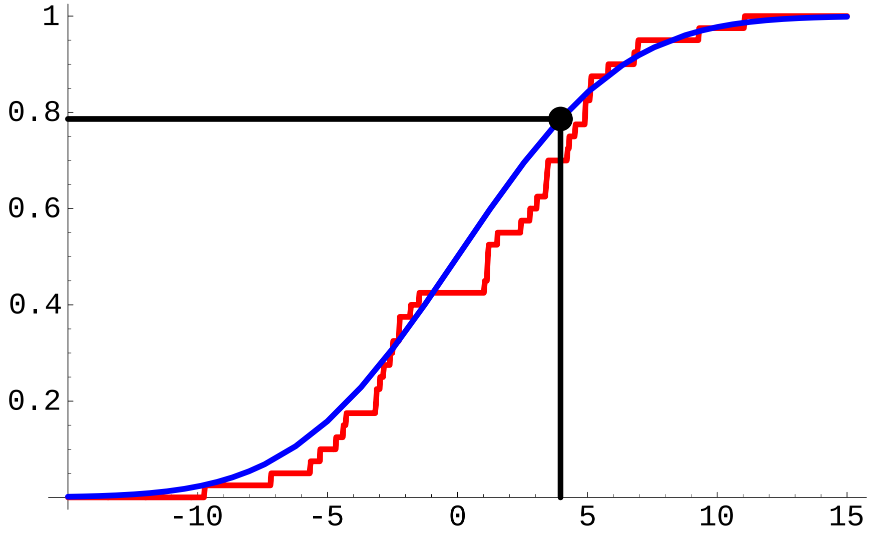
- Experimental, e.g. from many repeated measurements
- From a quality description. For example, if an observation has value 12 and st.dev = 2, its corresponding distribution is $\mathcal{N}(12, 2)$.

...assuming it has a normal distribution

Question.
What could be the parameters of the blue distribution in the figure?

To make a draw:

- Generate random number p between 0 and 1
- Idenitify a such that $F(a) = p$.



Remark. Compare Matlab command `randn`

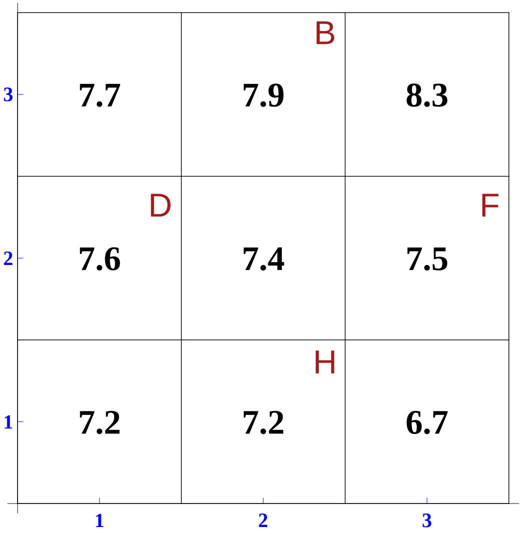
Example: DTM uncertainty

Suppose we are given the observations on the right.

Moreover, the st.dev of the observations is specified as $\sigma = 0.2$,

So,

$$B \sim \mathcal{N}(7.9, 0.2)$$
$$D \sim \mathcal{N}(7.6, 0.2)$$
$$F \sim \mathcal{N}(7.5, 0.2)$$
$$H \sim \mathcal{N}(7.2, 0.2)$$



	1	2	3
3	7.7	7.9	8.3
2	7.6	7.4	7.5
1	7.2	7.2	6.7

Monte Carlo simulation

One Monte Carlo experiment:

1. Draw k random numbers $p_{u_1}, \ldots, p_{u_k}$ between 0 and 1.
2. Determine the corresponding values $\tilde{u}_1, \ldots, \tilde{u}_k$, s.t. $F(p_{u_j}) = u_j$.
   Simulation of each of the observations
3. Determine $\tilde{v} = g(\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_k)$
   Simulation of the derived parameter

Repeat the experiment many (e.g. 10 000) times and collect the outcomes of each experiment

Gives 10 000 simulated values $\tilde{v}$.

Distribution of the different outcomes for $\tilde{v}$ gives insight on the sensitivity on the variation in the input observations u_i. (according to the given cdfs F_{u_i})

100 slope estimations

First run:

$$\hat{B}_1 = 7.86, \hat{D}_1 = 7.88, \hat{F}_1 = 7.47, \hat{H}_1 = 7.41,$$

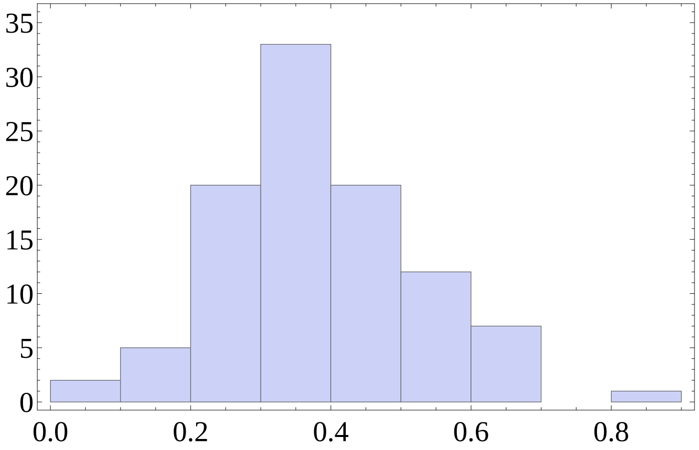
So

$$\hat{s}_1 = 0.31$$

Simulation consisting of 100 runs:

Mean slope: $\overline{s}_{100} = 0.382$;

St.dev. slope: $\sigma_{100} = 0.139$.



Results, Monte Carlo simulation

1st Conclusion.
We got a quality desciption of the slope estimation!

Method comparison
Could use MC as a tool to select between different estimation methods:
How do the simulation distributions compare?

In addition, full resulting distribution allows to

- Assess exceeding risks
- Assess the probability the slope is within a certain interval
- ...

Assumptions, Monte Carlo

Question
What assumptions did we make?

Question
What are the computational efforts?

Question*
Is the outcome also normally distributed?

Conclusions

The results of spatial data processing are largely useless if the quality is unknown.

Validating classification results is pretty straightforward using the confusion matrix.

Outliers may spoil your statistics:

- Evaluate your results: do they make sense?
- Could outliers play a role?
- Use robust methods like RANSAC if necessary

Beware:
Errors propagate from the observations in derived results.

Monte Carlo simulation is one technique to estimate the influence of errors.

Exercises

Exercise on RANSAC

Exercise 10.1 Consider a flat rock face sampled by 10 000 laser points. Approximately 1000 of the points are considered outliers. We are fitting a plane through the 3D points using RANSAC

a). How many points define a plane in 3D?

b). In what exceptional situation is it not possible to fix a plane with three different points?

c). What is the probability that a laser point belongs to the plane?

d). What is a good distance threshold? That is, what distance would you tolerate between a laser point and the fitted plane?

e). How many trials are needed to have a 90 % probability on at least one outlier free trial?

f). Same question for 99 %.

Exercise on Monte Carlo simulation

Exercise 10.2 In the course slides a slope is estimated from a 3×3 window based on the four closest neighbours.

a). Sketch how a slope can be estimated in a similar way from a all eight neighbours.

b). Do you expect that the simulated Monte Carlo st.dev will increase or decrease if we use more observations in this way (assuming all observations have the same quality). Why?

An alternative approach is to first fit a plane through all nine observations (compare Lecture 6) and determine the slope from the plane.

c). Give the formula for the slope of a plane given by the equation $z = ax + by + c$.

d). Determine the slope of the plane best fitting the observations of Exercise 6.8.

e). Sketch how to use the Monte Carlo framework to derive a st.dev. of the slope estimation using the fitted least squares plane.

f). Perform the simulation in e.g. Matlab (100 runs) and compare the outcomes with the outcomes in the slides (using the four closest neighbours). Use the observations given in Exercise 6.8 and take $\sigma = 0.2$ for the quality of the observations.

Answers, Exercise 10.1

a) In general three points define a plane.

b) When the three points are all on one line

c) In this case 90 %

d) Something like 5 cm (compare to specifications of scanner devices)

e) The probability on an error free trial. To get 90% probability: 2 trials is enough; (Compare slide 23, here m=3; p=0.9; z=.9).

f) 99% probability: 4 trials is enough. Use now z=99.

Answers, Exercise 10.2

a) Compare slide 25. you could estimate the gradient in the x-direction by taking the average of 0.5(C-A), 0.5(F-E) and 0.5(I-G). Similarly for the gradient in the y-direction.

b) I expect the st.dev will become smaller. More observations used, so planes wil be more similar, and less varying in slope.

c) Compare, Lecture 9, slide 13. slope = $\sqrt(f_x^2 + f_y^2)$. Write $f = ax + by + c$; Then $fx = a$; $fy = b$; so the slope equals $\sqrt(a^2 + b^2)$

d) Exercise 6.8: plane parameters are: a=0; b = 0.47; c = 6.57; So the slope equals 0.47 as well.

e) In each run, a new vector of observations vecy_sim is generated, consisting of the original observations of Exercise 6.8 with random noise added. To generate the noise, a normal distribution with a st.dev of 0.2 is used. So, one such simulated vector of observations could look like

vecy_sim = vecy + vec_noise =

(7.2, 7.2 ,6.7, 7.6, 7.4, 7.5 ,7.7 ,7.9 ,8.3) + (-.12, 0.03, 0.05, -0.12, .06, -.05, .07, .1, -.01)

Then the least squares plane is fitted like in Exercise **5.1 but using vecy_sim instead of vecy. This results in

xhat_sim = (a_sim, b_sim, c_sim).

Using a_sim and b_sim the slope s_sim is determined as $s_{sim} = \sqrt(a_{sim}^2 + b_{sim}^2)$. This is the result of one run. Many runs (e.g. 100) result in a variety of slopes that can be used to estimate a standard deviation.

f) This is left to the student.