Characterisation of surface roughness

David Nečas and Petr Klapetek

Masaryk University, Brno, Czech Republic
yeti@physics.muni.cz

Greifswald, 25 June 2012

Outline

Roughness

Roughness of solid surfaces is a common phenomenon.

Quantitiative characterisation is important for

- technological process optimisation
- deeper understanding of phenomena behind surface treatment
- interpretation of measurement in
  - nanoindentation
  - optics – ellipsometry, spectrophotometry, ...
  - X-ray photoelectron spectroscopy
  - ...
- surface chemistry, also biology

Roughness can be characterised using the above methods. They are sensitive to different properties and applicable to different kinds of rough surfaces.

Atomic force microscopy AFM (or profilometry, at a larger scale) provides directly the **topography** of the surface – the ultimate information.

Rough surface near Cinder Cone, California, USA

Rough surface of ZnSe film – AFM data visualisation

Atomic force microscopy data

AFM provides heightfields of limited

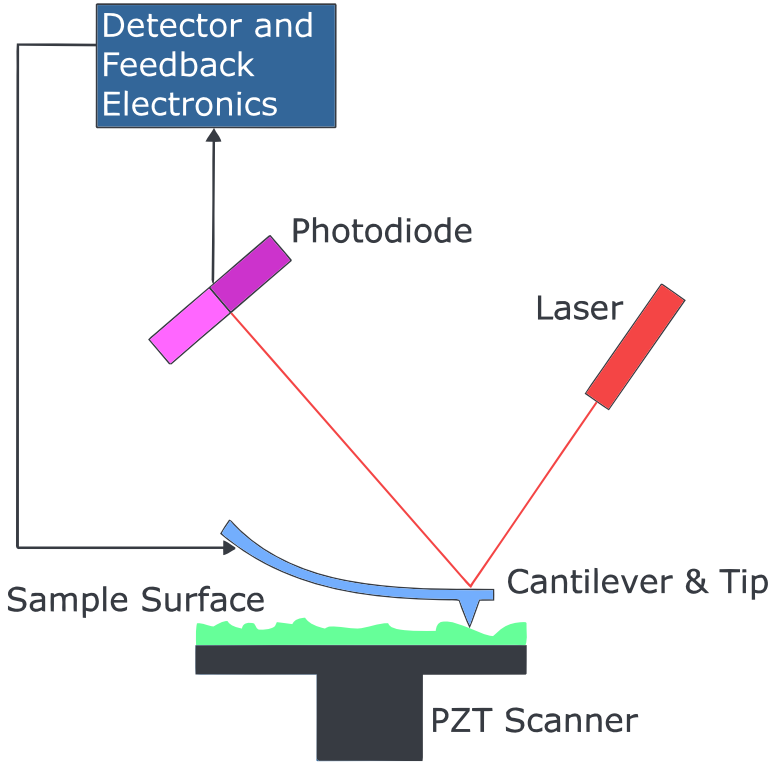
- size
- resolution (often due to tip concolution)
- maximum slope (due to tip concolution)

Too fine, too coarse or too steep features are not adequately captured.

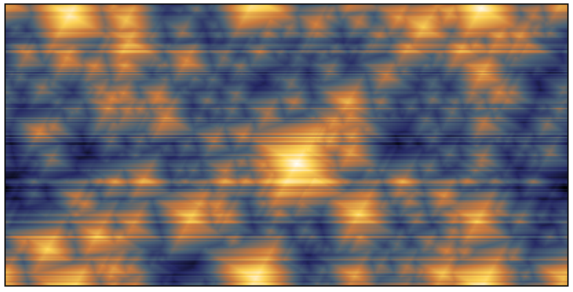
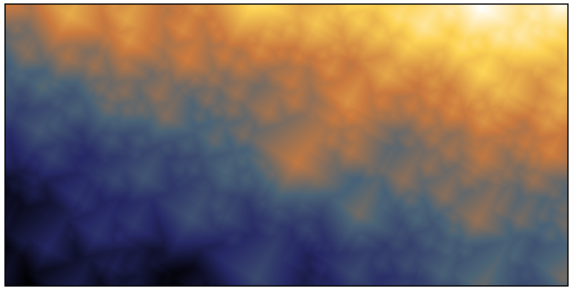
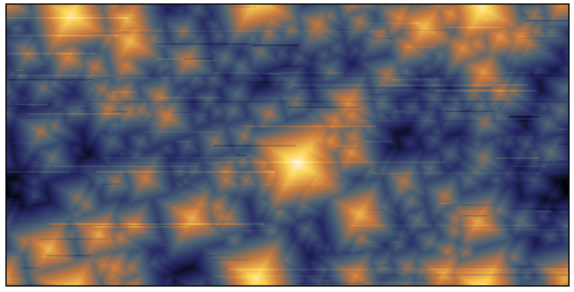
The data rarely can be directly analysed. Usually, **corrections** are necessary:

- levelling (plane, polynomial)
- line-by-line-match
- local defect correction
- Fourier-domain filtering

All represent serious data manipulation.



Atomic force microscope scheme

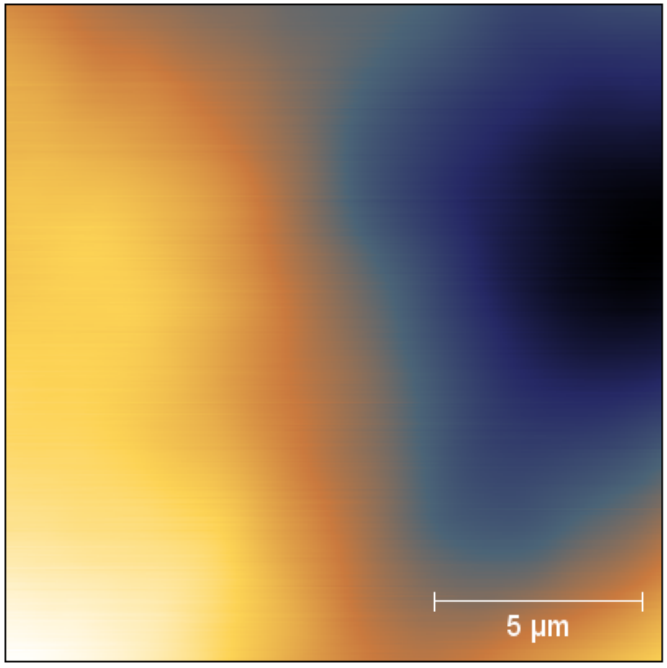




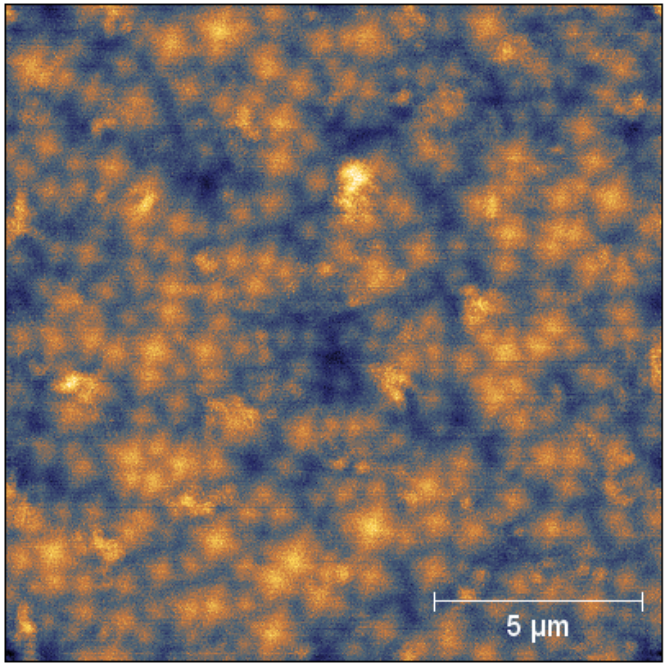
Visual assessment

A widespread practice is to just publish AFM images.

Because, of course, everyone can easily tell from a topographical image how rough or smooth the surface is.

For instance, the left surface is obviously smooth while the right one is rough:



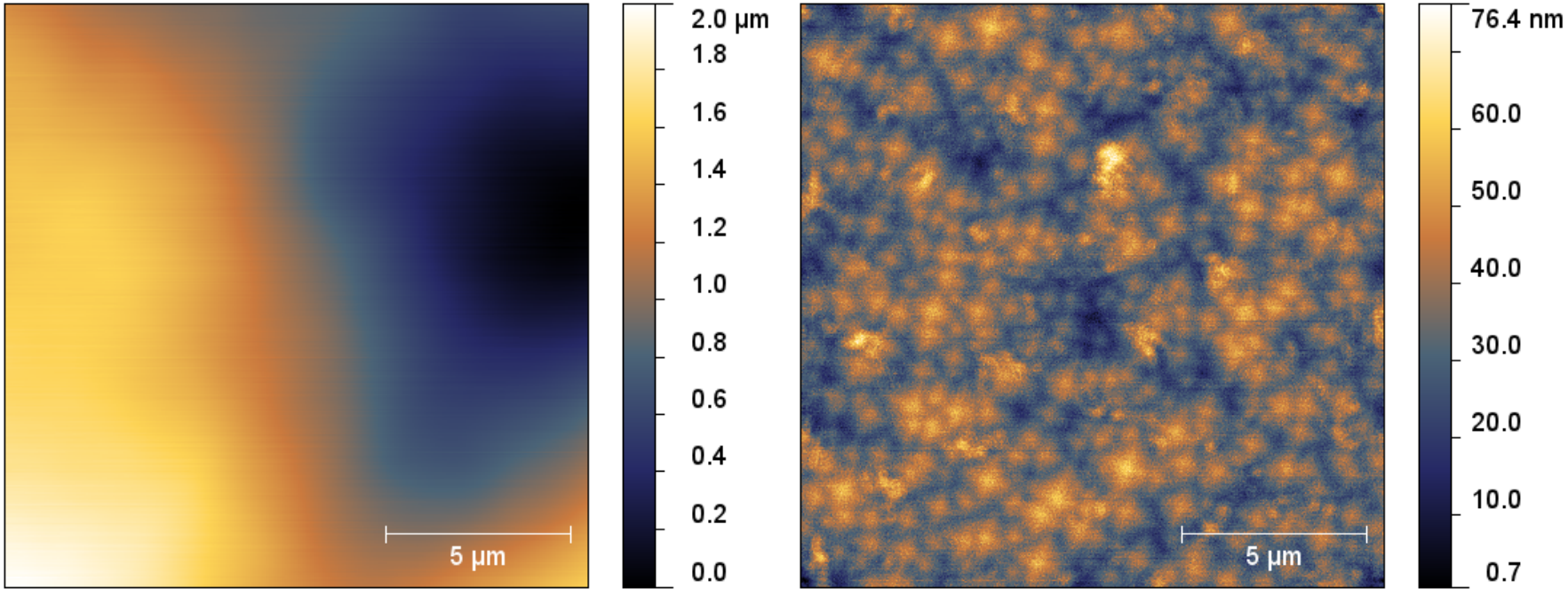


Visual assessment

A widespread practice is to just publish AFM images.

Because, of course, everyone can easily tell from a topographical image how rough or smooth the surface is.

For instance, the left surface is obviously smooth while the right one is rough:



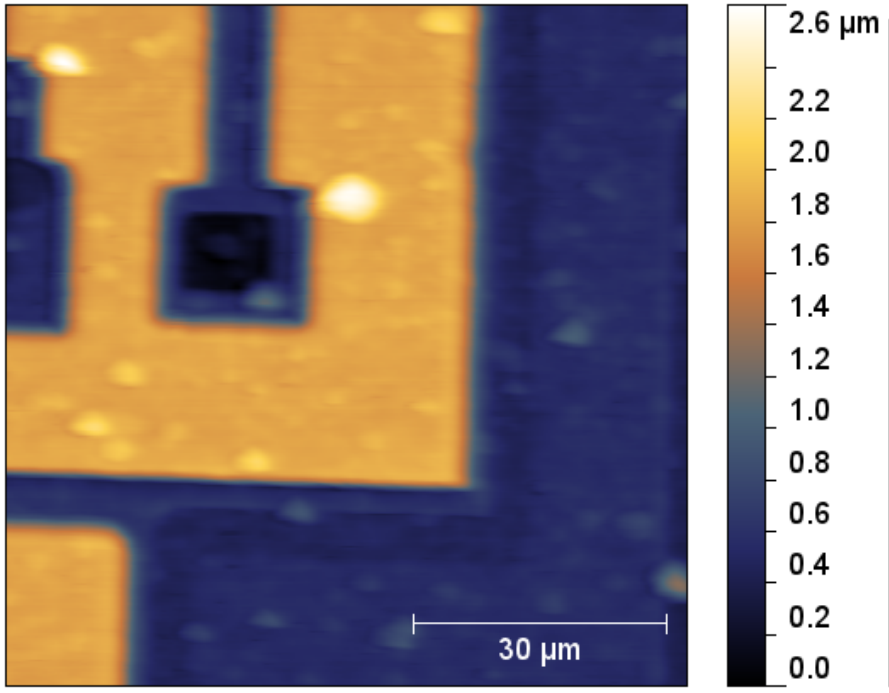
Oops! We forgot the scales (also common).

But wait, now it is no longer clear which is which.

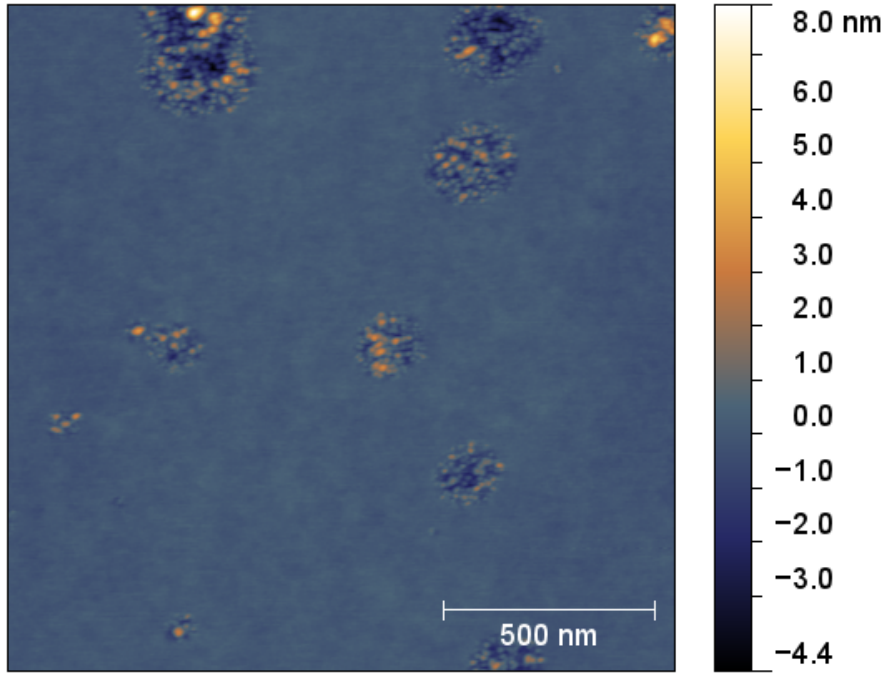
More difficulties

It can get even more complicated.

Large topographic features can be mixed with roughness in various ways.



Microchip surface



TiO_2 film at a certain growth stage

Maybe we have to resort to mathematics after all.

And consider the simple case first: Surfaces corresponding to wide-sense stationary stochatic random processes. In other words, they are **the same everywhere** from the statistical point of view.

Quantification

To goal is to characterise the roughness, i.e. to represent the information from the image using just a **small number of quantities**.

Balance is needed between not being able to compare because we have too much data and because we have reduced the data too much.

- scalar – a single number (or set of them)
- one-dimensional characteristic – distributions, correlation functions
- two-dimensional – rare, usually an intermediate step

Approaches:

- Use a **norm**: ASME B46.1-1995, ASME B46.1-1985, ISO 4287-1997, ISO 4287/1-1997, . . . This is the safe path to comparable results – sort of. But we may not have any idea how is *the height of the third highest peak from the third lowest valley per sampling length* related to anything we are trying study.

  Most norms have origin in profilometry – i.e. for profiles, defining 1D parameters called $R_{\text{something}}$. 2D are called $S_{\text{something}}$.
- Try to calculate quantities that are **relevant** for other measurements or processes – for instance power spectrum density for optics. This may be is difficult and even more difficult to compare.

Several basic concepts and quantities are common to both approaches.

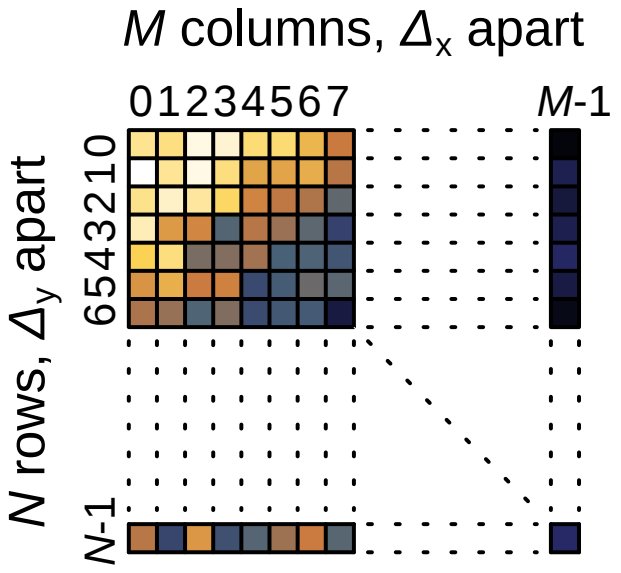
Sampling

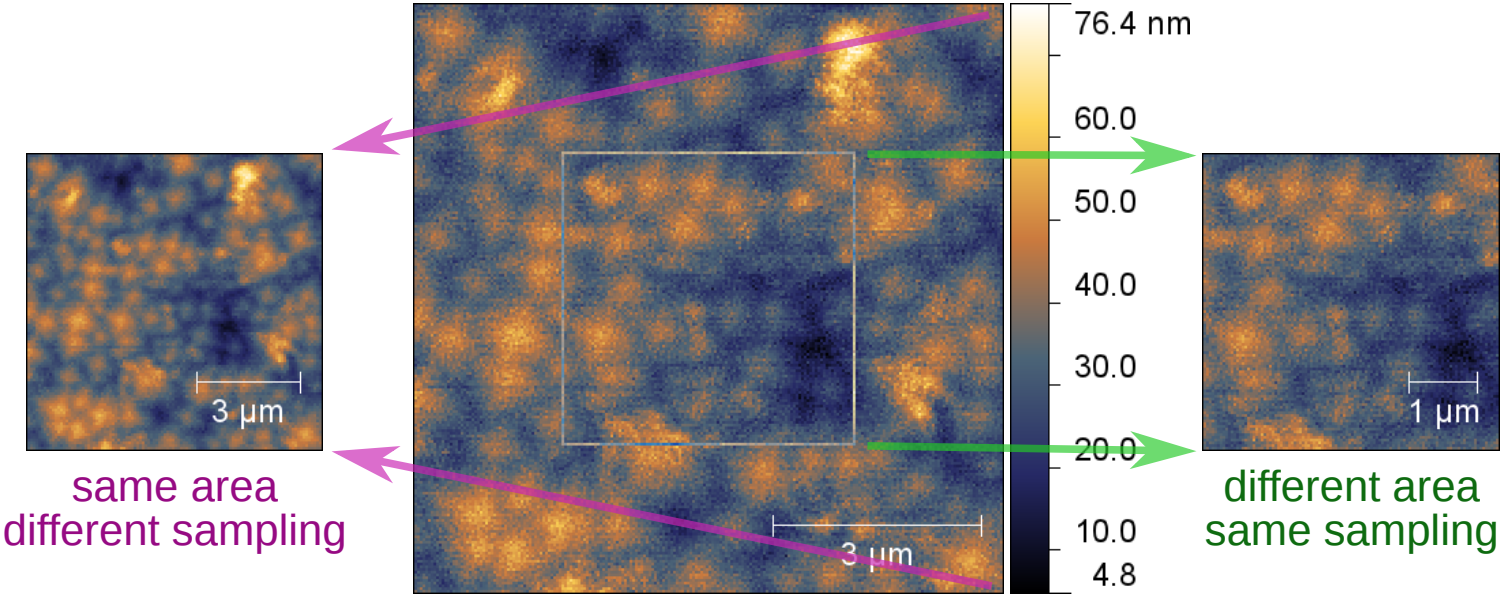
The surface can be imagined as a two-dimensional function $z(x, y)$ defined on entire $\mathbb{R} \times \mathbb{R}$.

Measurement **samples** this function at discrete points in a finite area.

Thus we obtain a matrix of heights $z_{i,j}$, where $i = 0, 1, 2, \ldots, N - 1$ and $j = 0, 1, 2, \ldots, M - 1$.

But we try to find properties of the surface $\Rightarrow$ good characteristics should be **sampling-independent**.





Height distribution characteristics

The simplest characteristics depend on the **distribution of heights** $\varrho(z)$ and measure the roughness magnitude.

Assumption: The mean value is zero (possible since the surface is the same everywhere)

$$\sum_{i=0}^{N-1}\sum_{j=0}^{M-1} z_{i,j} = 0 \qquad \text{that approximates} \qquad \lim_{\Omega\to\mathbb{R}^2} \frac{1}{|\Omega|}\int_\Omega z(x,y)\,\mathrm{d}x\,\mathrm{d}y = 0$$

Root mean square roughness

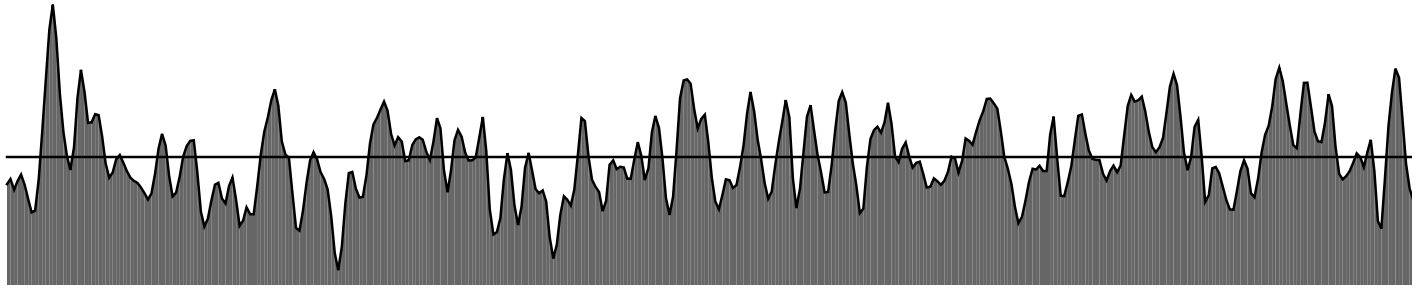
$$S_\mathrm{q} = \sigma = \left[\frac{1}{MN}\sum_{i=0}^{N-1}\sum_{j=0}^{M-1} z_{i,j}^2\right]^{1/2}$$

$$\sigma^2 = \int z^2\varrho(z)\,\mathrm{d}z$$

Average roughness

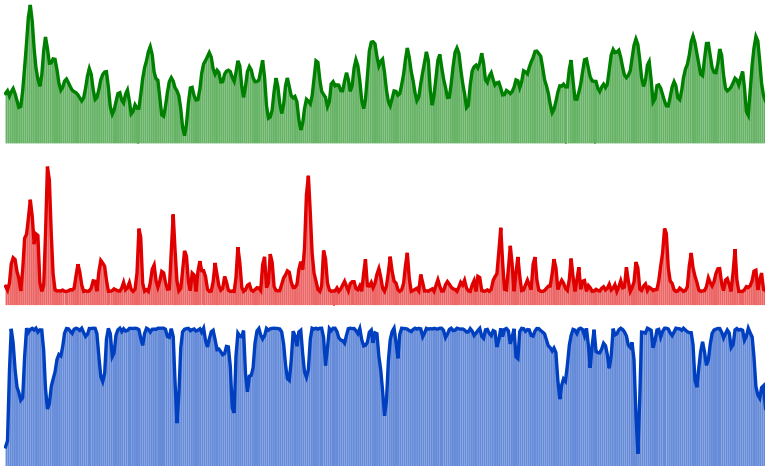
$$S_\mathrm{a} = \frac{1}{MN}\sum_{i=0}^{N-1}\sum_{j=0}^{M-1} |z_{i,j}|$$

$$S_\mathrm{a} = \int |z|\varrho(z)\,\mathrm{d}z$$



Sampling independent – mostly.

Beyond S_a and S_q



These profiles were taken from surfaces with quite similar both S_a and S_q.

But the roughness is obviously different.

And in certain applications the surfaces would behave quite differently.

A number of other height distribution parameters are aimed at capturing this.

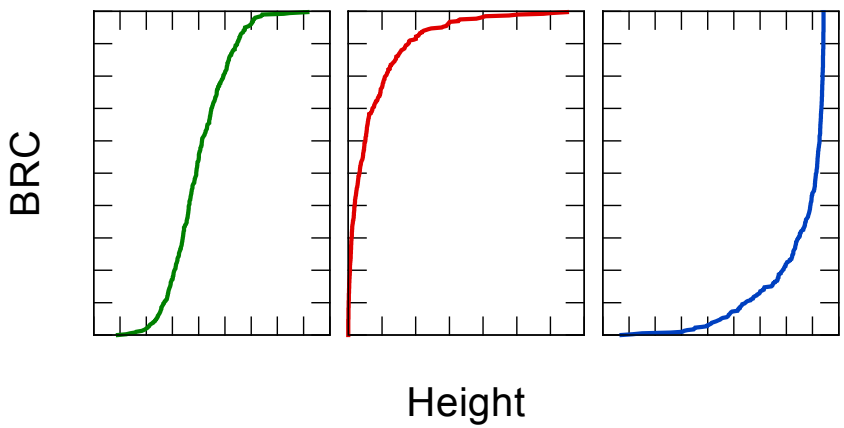
Peak and valley parameters – or minimum and maximum in plain terms

$$S_p = \max_{i,j} z_{i,j} \,, \qquad S_v = \min_{i,j} z_{i,j}$$

Useful for indication of unusual sharp spikes or cracks.

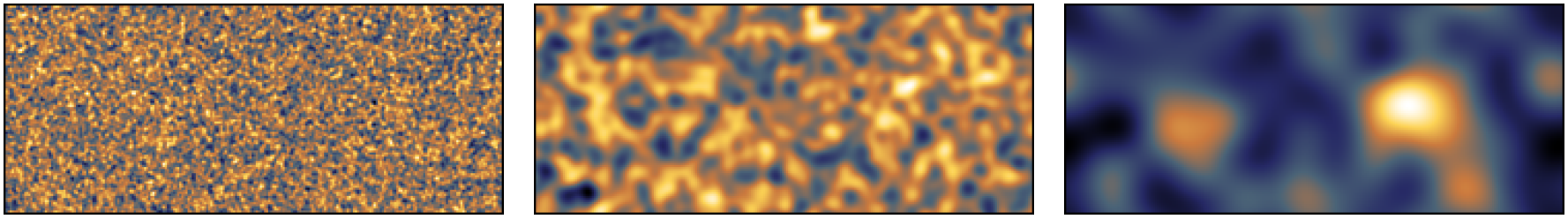
Skewness S_{sk}, **kurtosis** S_{ku} – third and fourth order moments of $\varrho(z)$.

One-dimensional characteristics: **bearing ratio curve** (BRC). In plain terms, the cummulative height distribution.



Spatial characteristics

Surfaces can have the same height distributions yet differ a lot:

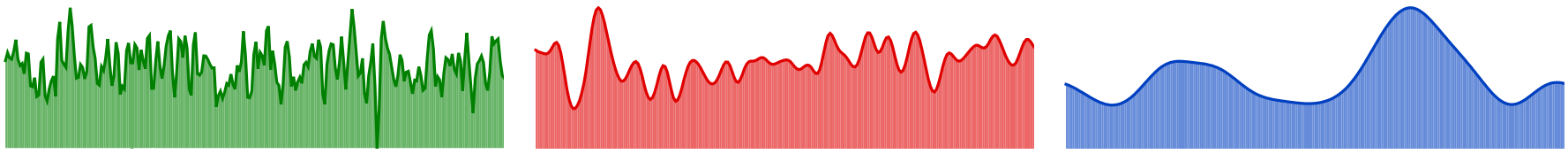


The difference is in **spatial properties**. Listing only S_q as the complete roughness description is wrong.

In AFM we often determine spatial properties along **scan lines**, i.e. profiles along the fast scanning axis:

- more reliable – does not rely on correct mutual alignment of lines
- simpler – spatial properties in two dimensions can be complex

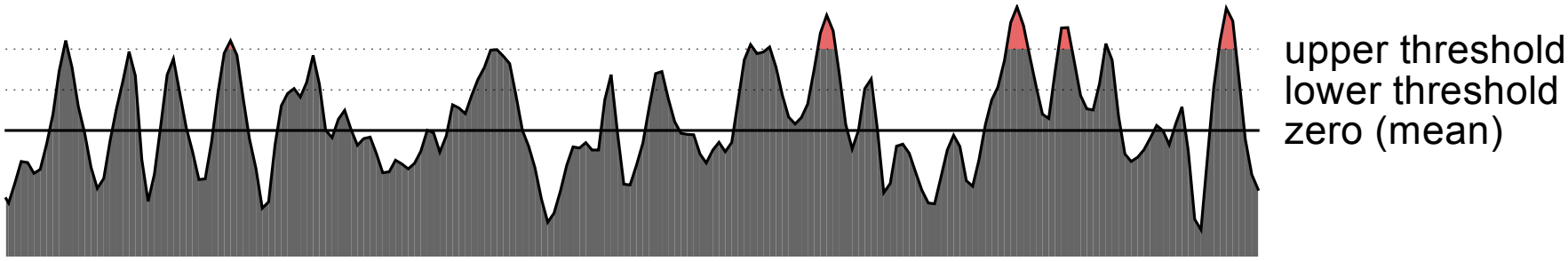
Hence we use characteristics of one-dimensional data for two dimensional-data – possibly with futher processing over individual lines (accumulation, averaging).



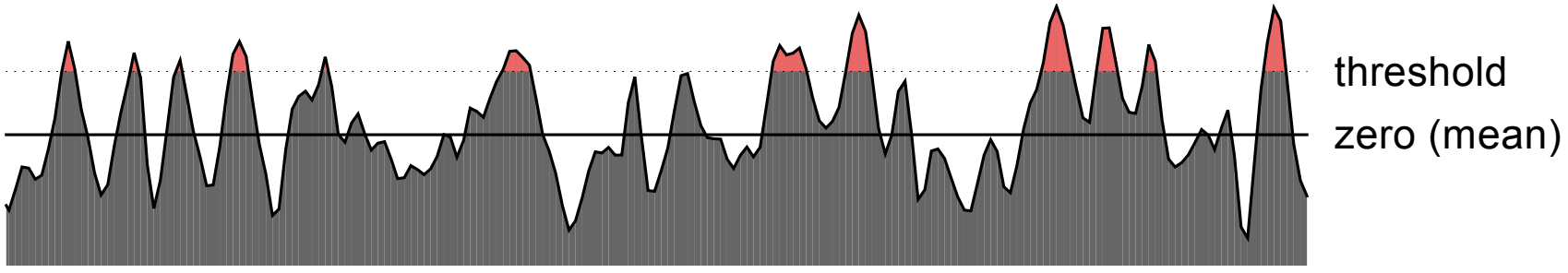
This is all right if the surface is **isotropic**.

Peak counting

Norms usually employ **peak counting** and require to choose certain **thresholds**.



Peak *must extend above the upper threshold and then fall below the lower threshold to be counted.*



High spot *is simpler – defined by crossing only one threshold.*

Peak count P_c – number of peaks per unit length.

High spot count HSC – number of high spots per unit length.

Mean spacing S_m – average distance between zero crossings (in one direction).

Similarly for valleys.

Hybrid parameters

Another class of parameters is related to local **slopes**.

Local derivatives are approximated as $(z_j - z_{j-1})/\Delta_x$ – somewhat crude and sensitive to sampling.

Average slope

$$\Delta_\mathrm{a} = \frac{1}{N-1}\sum_{j=1}^{N-1}\frac{|z_j - z_{j-1}|}{\Delta_x} \qquad \text{approximates} \qquad \Delta_\mathrm{a} = \lim_{L\to\infty}\frac{1}{L}\int_0^L |z'(x)|\,\mathrm{d}x$$

Root mean square average slope

$$\Delta_\mathrm{q} = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N-1}\left[\frac{z_j - z_{j-1}}{\Delta_x}\right]^2} \qquad \text{approximates} \qquad \Delta_\mathrm{q} = \lim_{L\to\infty}\sqrt{\frac{1}{L}\int_0^L \left(z'(x)\right)^2\,\mathrm{d}x}$$

Actual (developed) profile length

$$L_0 = \sum_{j=1}^{N-1}\sqrt{\Delta_x^2 + (z_j - z_{j-1})^2} \qquad \text{approximates} \qquad L_0 = \lim_{L\to\infty}\int_0^L \sqrt{1 + \left(z'(x)\right)^2}\,\mathrm{d}x$$

Autocorrelation function

Influence of roughness on physical quantities depends, however, on **correlation** properties and/or **spatial frequencies** present in the surface.

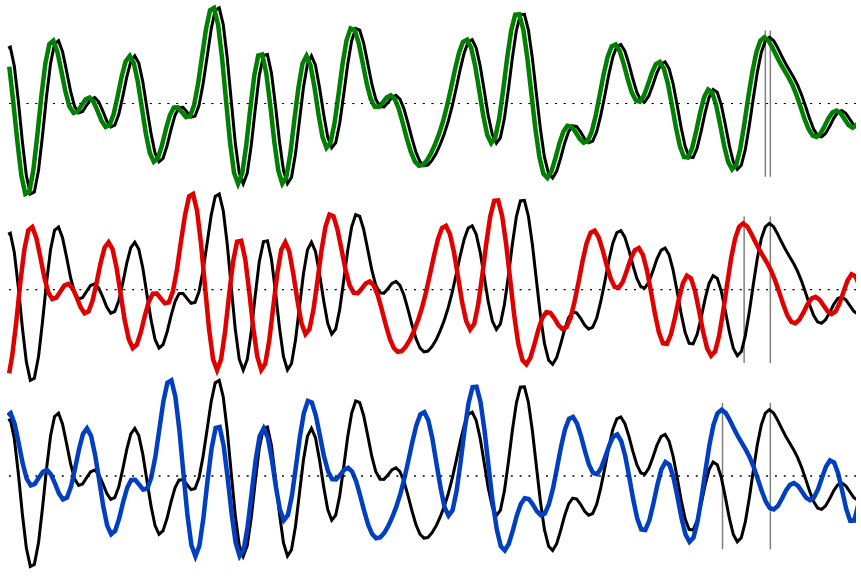
Autocorrelation function (1D)

$$G_x(\tau_x) = \lim_{L\to\infty} \frac{1}{2L} \int_{-L}^{L} z(x)\, z(x+\tau_x)\, \mathrm{d}x$$

approximated for $\tau_x = k\Delta_x$

$$G_k = \frac{1}{M-k} \sum_{j=0}^{M-1-k} z_j z_{j+k}$$

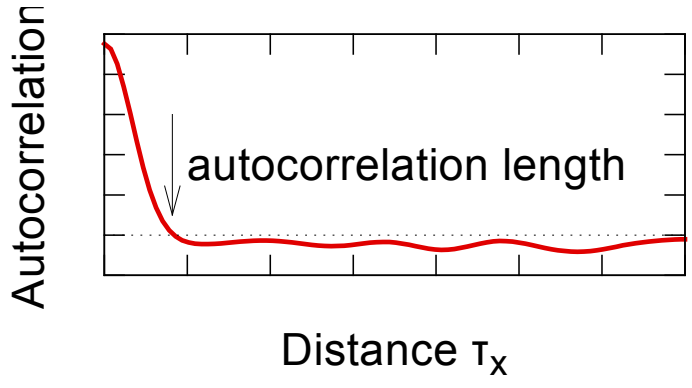
usually with averaging over rows.



ACF expresses how the surface is correlated (similar) to itself at distance τ_x.

It goes to zero for τ_x so large that the heights become independent.

This leads to the notion of **autocorrelation length** (or autocorrelation distance) T.



Modelling

Gaussian

$$G_x(\tau_x) = \sigma^2 \exp\left(-\frac{\tau_x^2}{T^2}\right)$$

Common model since Gaussian has many nice properties that simplify calculations.

Exponential

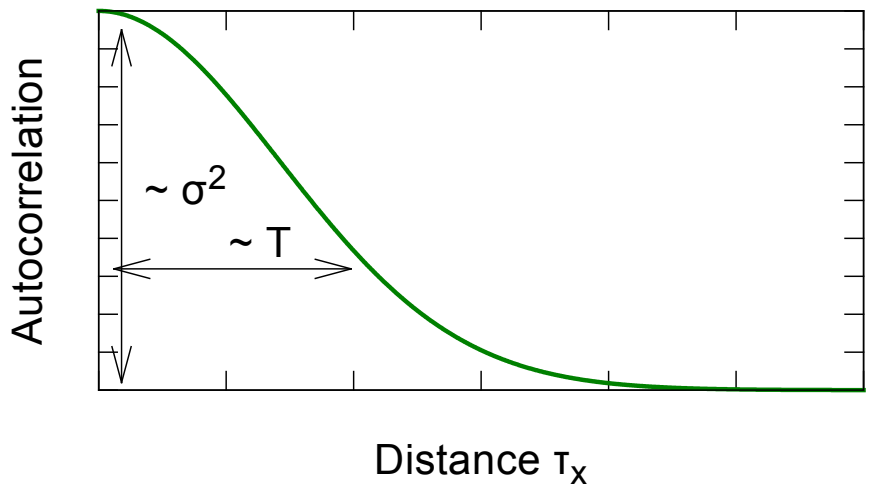
$$G_x(\tau_x) = \sigma^2 \exp\left(-\frac{\tau_x}{T}\right)$$

Arguably more likely to be the result of certain physical proceses.

How do we actually get the root mean square roughness σ and autocorrelation length T?

Using **fitting**:

- calculate discretised G_k from experimental data (an efficient algorithm is FFT-based)
- select a model for $G_x(\tau_x)$ – usually one of the two above
- fit the model on the discretised ACF



This works well if G_k converges to $G_x(\tau_x)$ as the measurement area gets larger.

It does (pointwise) but with some quirks.

Discrete versus continuous

Fun fact: If the surface mean value is zero then

$$\frac{M}{2} G_0 + \sum_{k=1}^{M-1} (M-k) G_k = 0 \,.$$

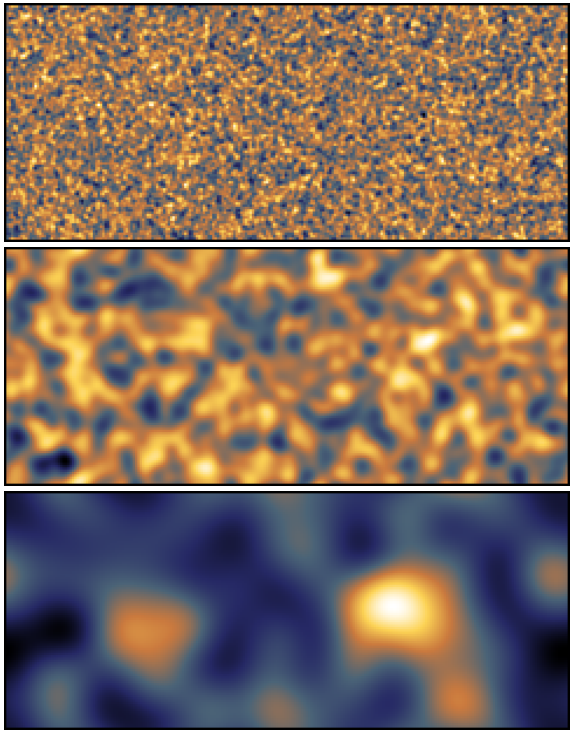
This means G_k must take **both signs** (or $G_k \equiv 0$).

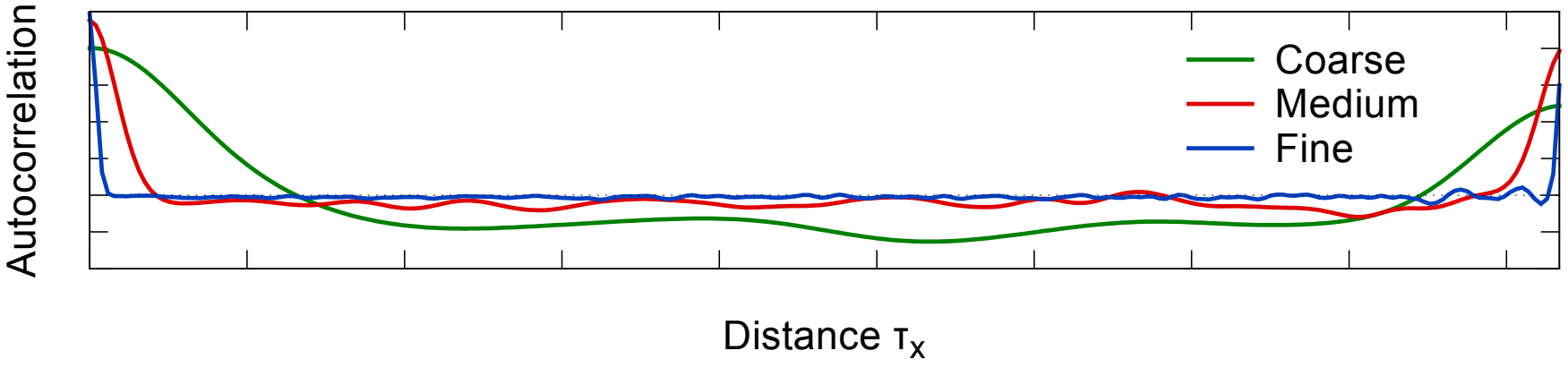
But the model ACF is **nonnegative** everywhere.

How is that possible?

Discretised G_k for large k is calculated from only a few values – not converged, unreliable.

We can throw away the worthless part or use weighting.





Power spectrum density

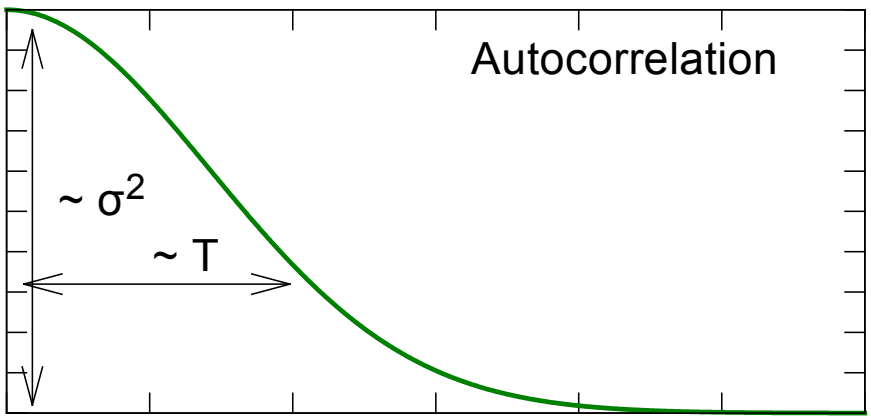
Spatial relations are often expressed using another quantity: the **power spectrum density** function.

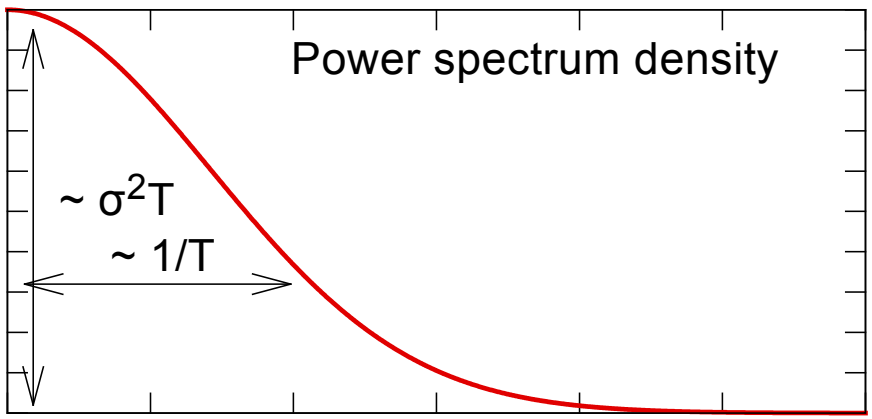
It describes the content of **spatial frequencies** in the surface. It is the modulus of **Fourier transform**:

$$W_x(k_x) \propto \lim_{L\to\infty} \frac{1}{2L} \left| \int_{-L}^{L} z(x) \mathrm{e}^{\mathrm{i}k_x x}\, \mathrm{d}x \right|^2$$

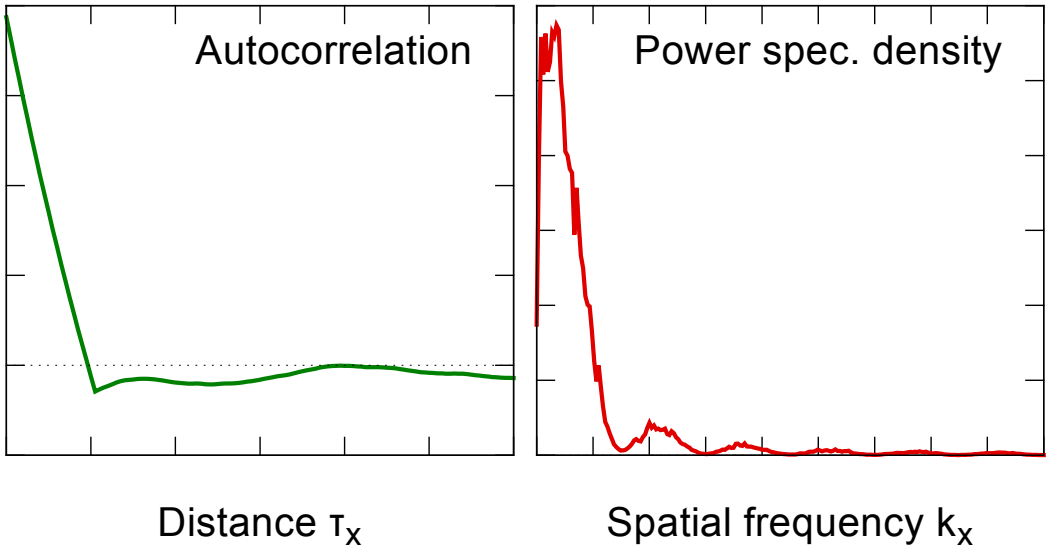
Due to Wiener–Khinchin theorem for stationary process $W_x(k_x)$ and $G_x(\tau_x)$ are Fourier transforms of each other.

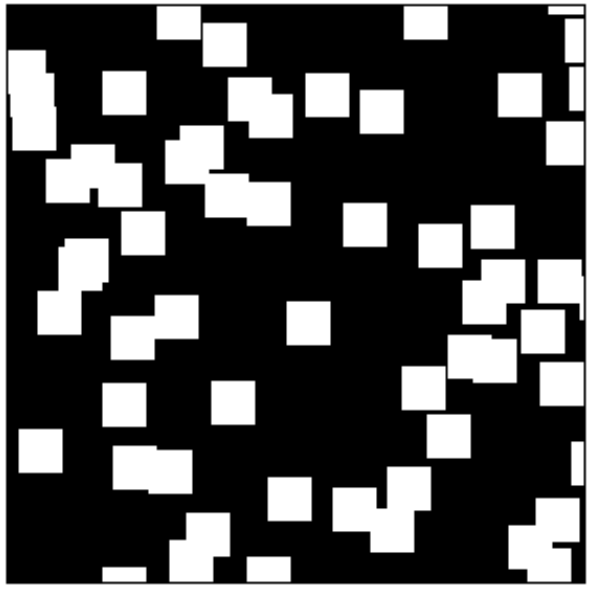
So they bear the same information but expressed differently.

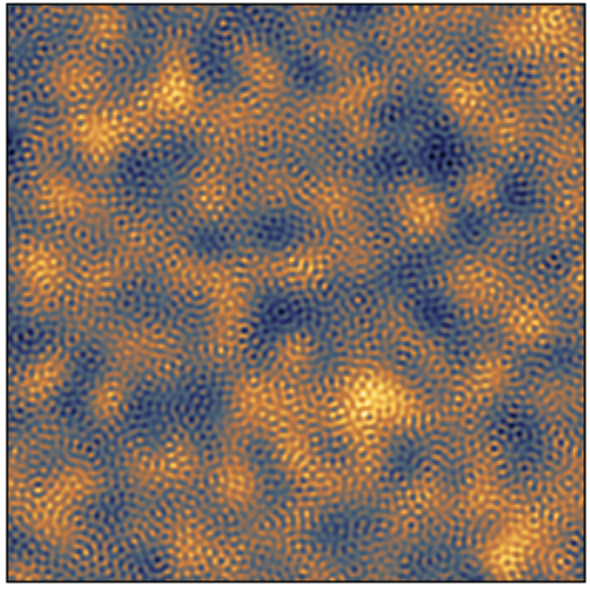




PSDF versus ACF







Modelling

Gaussian (corresponds to Gaussian ACF)

$$W_x(k_x) = \frac{\sigma^2 T}{2\sqrt{\pi}} \exp\left(-\frac{k_x^2 T^2}{4}\right)$$

Exponential (corresponds to exponential ACF)

$$W_x(k_x) = \frac{\sigma^2 T}{\pi} \frac{1}{1 + k_x^2 T^2}$$

Fitting of these model functions is again used to obtain T and σ.

Discretised PSDF is calculated using **discrete Fourier transform** of the data:

$$W_\nu \propto |Z_\nu|^2 , \qquad \text{where} \qquad Z_\nu = \frac{1}{\sqrt{M}} \sum_{j=0}^{M-1} z_j \mathrm{e}^{-2\pi \mathrm{i}\, j\nu}$$

We only write $\propto$ because no generally agreed-upon convention exists:

- continuous FT exponent can be either $\mathrm{i}k_x x$ or $2\pi\mathrm{i}k_x x$ – circular vs. straight frequency,
- direct FT followed by inverse FT must give the original function but this still permits an arbitrary split of multiplicative factors between the direct and inverse FT,
- we cannot prescribe a specific normalisation of both the Fourier coefficients and spectrum density,
- discretisation of normalisation for finite energy vs. finite average power functions.

The result: **a mess**. PSDF from different software is difficult to compare.

Energy theorem

Are some PSDF conventions better than others?

We impose: **Plancherel/Parseval theorem** – sum/integral of power spectrum is the same as sum/integral of squared data values.

There is no shortage of them:

Discrete Fourier transform

$$\sum_{j=0}^{M-1} z_j^2 = \sum_{\nu=0}^{M-1} W_\nu$$

Classic Fourier series of a periodic function

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} z(x)^2 \,\mathrm{d}x = \sum_{n=-\infty}^{\infty} |a_n|^2$$

Continuous Fourier transform

$$\int_{-\infty}^{\infty} |z(x)|^2 \,\mathrm{d}x = \int_{-\infty}^{\infty} W(k_x) \,\mathrm{d}k$$

Continuous Fourier transform of finite average power functions

$$\lim_{L\to\infty} \frac{1}{2L} \int_{-L}^{L} |z(x)|^2 \,\mathrm{d}x = \int_{-\infty}^{\infty} W_x(k_x) \,\mathrm{d}k$$

But then comes the **sampling-independence** requirement (recall it?): PSDF should be independent of how large part of the surface we measure and the sampling step.

Then the discrete PSDF must be just **sampled continous** $W_x(k_x)$. No additional factors.

Sampling-independence

Discretised PSDF correctly sampling the continuous one (circular frequency) is thus:

$$W_\nu = \frac{\Delta_x}{2\pi} |Z_\nu|^2 \ , \quad Z_\nu = \frac{1}{\sqrt{M}} \sum_{j=0}^{M-1} z_j \mathrm{e}^{-2\pi \mathrm{i}\, j\nu}$$

Units of W_ν are [dimension][height^2].

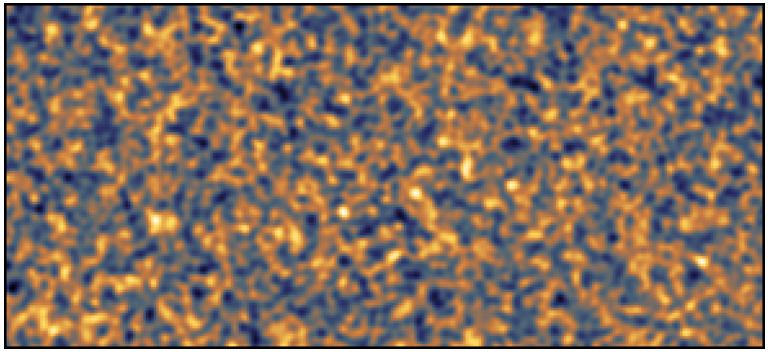
Units can be used to distinguish Fourier transform conventions:

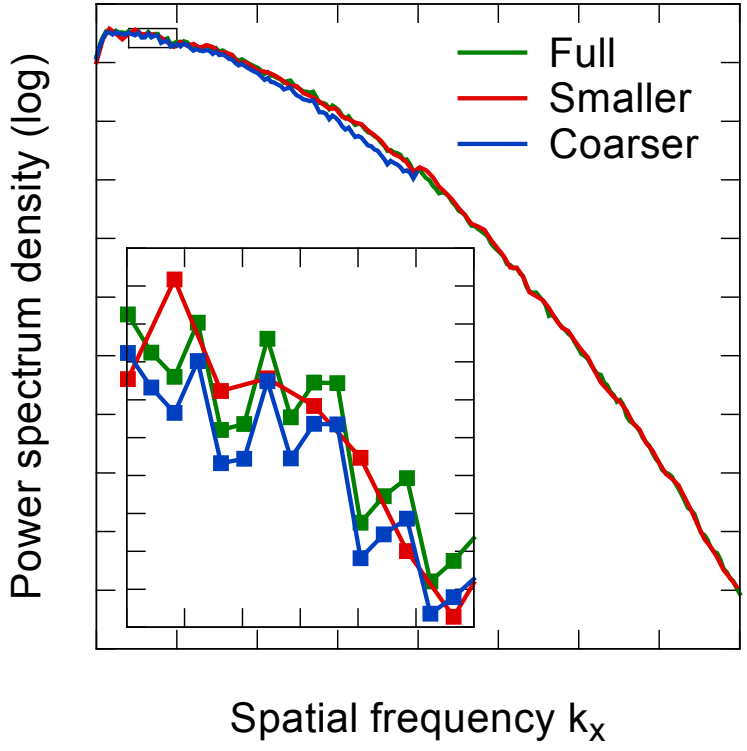
- [height^2] – plain DFT
- [dimension^2][height^2] – finite-energy continous FT

Data sampling change effect:

- coarser data – smaller max. frequency
- smaller data – coarser PSDF sampling

Change of variables, e.g. circular versus non-circular. Use that PSDF is **density**: $W_x(a)\,\mathrm{d}a = W_x(b)\,\mathrm{d}b$.





Windowing

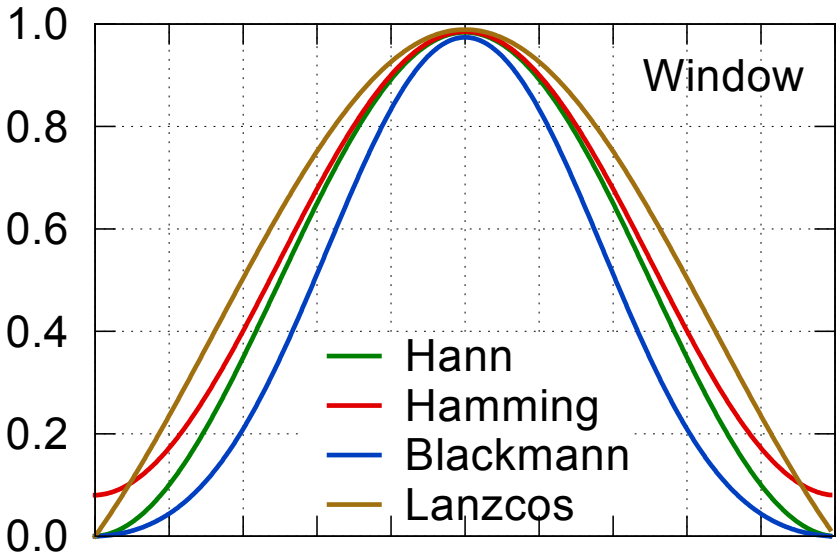
Fourier coefficients of data contatining a **discontinuity** decrease slowly with frequency – generally as $1/\nu$. This includes abrupt value change between the **right and left edge**.

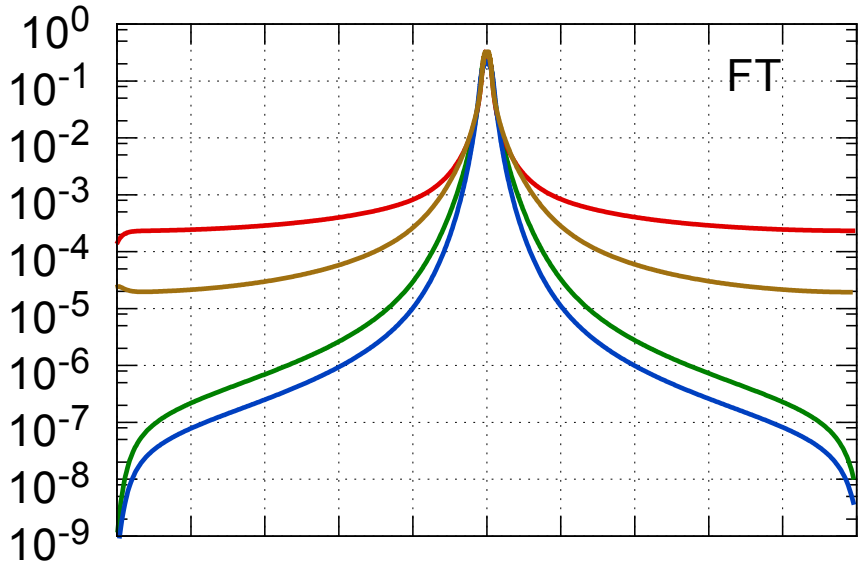
Most of real surfaces is **not periodic** with sampling length being an integral multiple of the period. This would completely **ruin** the PSDF estimate.

Solution: **windowing** – multiplying data with a smooth window function that falls to zero at the edges.

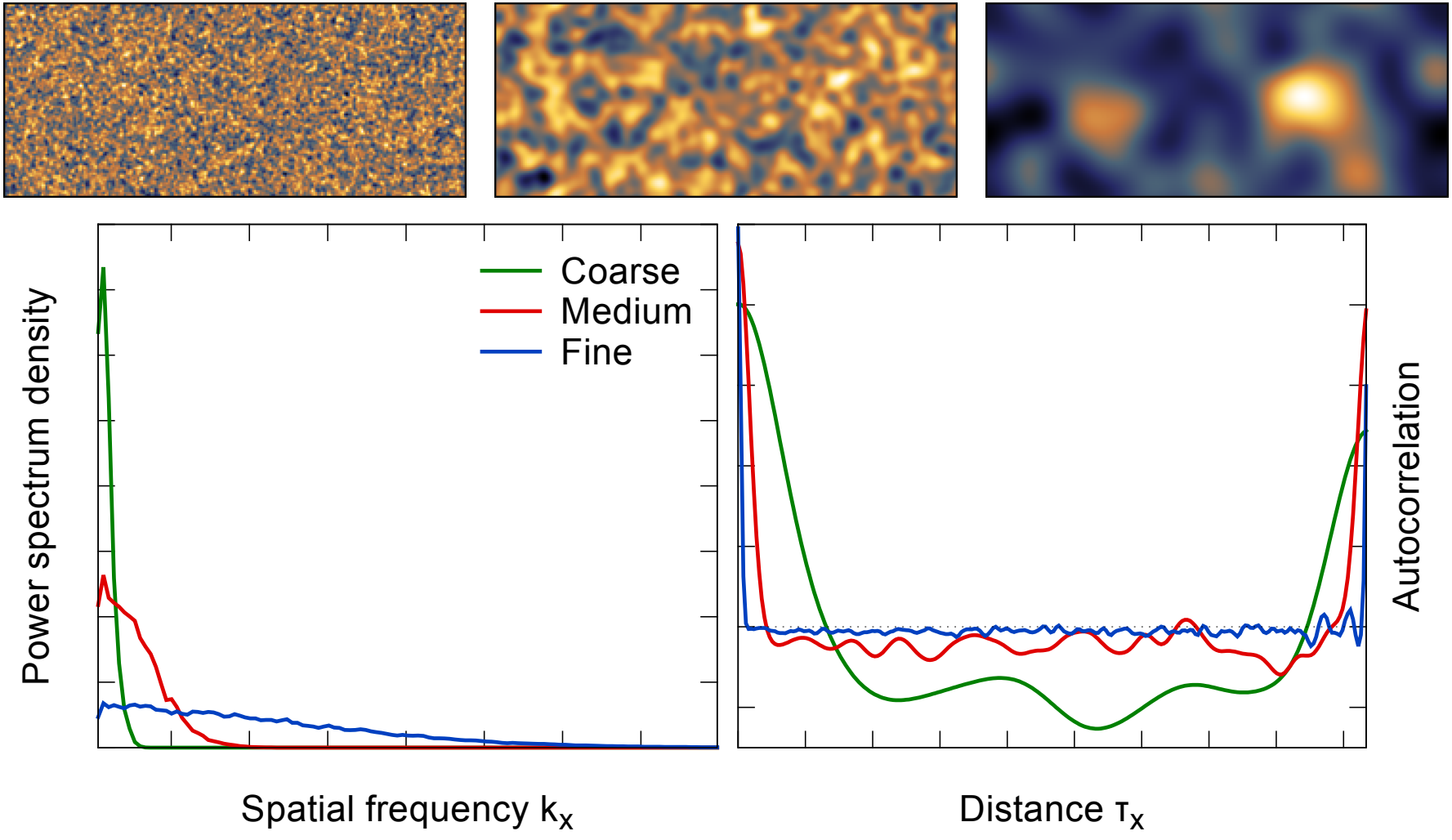
FT of data is **convolved** with FT of window function – spectral leakage.

Furthermore, σ is changed by windowing – needs correction, for instance to ensure that the RMS of windowed data is the same as original.





Examples



Since all three surfaces have the same σ area under all three PSDF curves is the same.

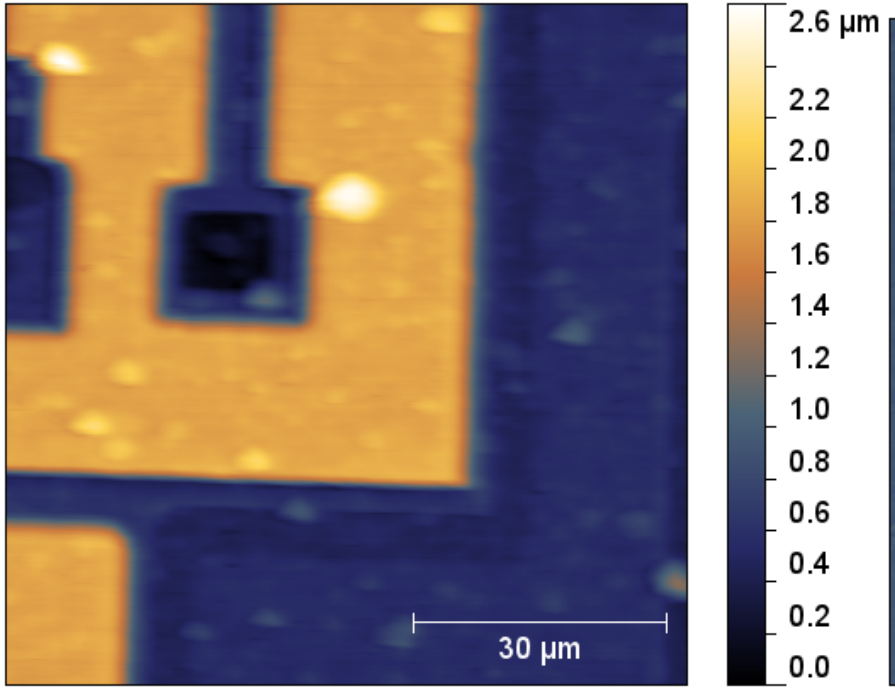
Non-uniform surfaces

Some surfaces are **not the same** everywhere:

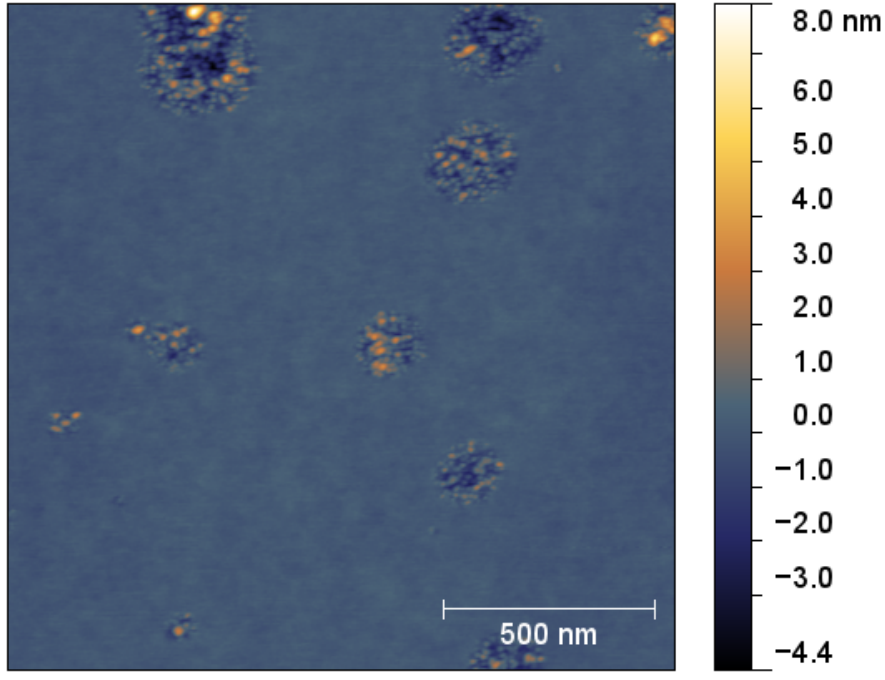
- **overall topgraphy** needs to be separated from roughness
- **non-uniform roughness**, i.e. different statistical properties at different locations

We are on thin ice evaluating them – getting comparable results is difficult.

Most methods require tuning of some parameters or making other choices based a user discretion.

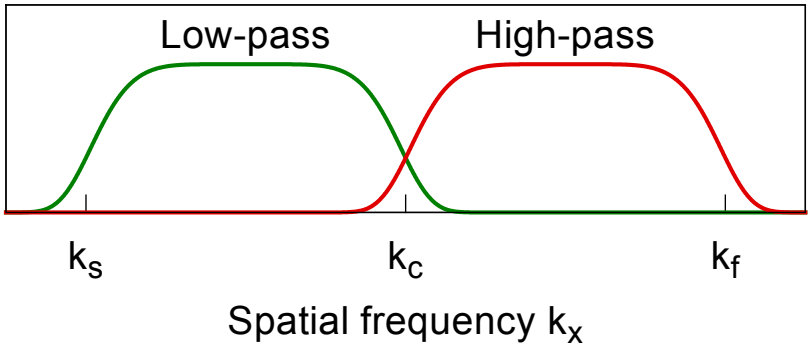


Microchip surface



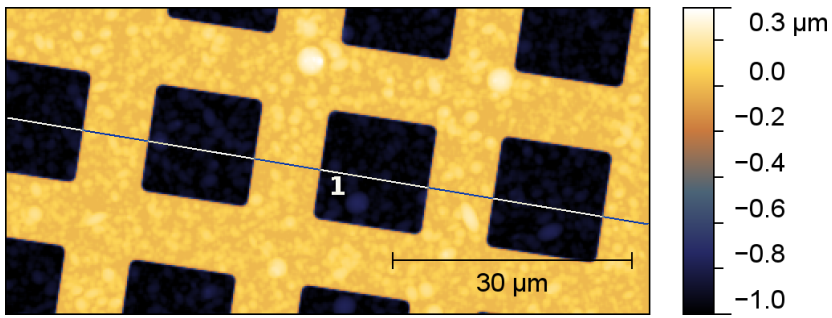
TiO_2 film at a certain growth stage

Overall topography

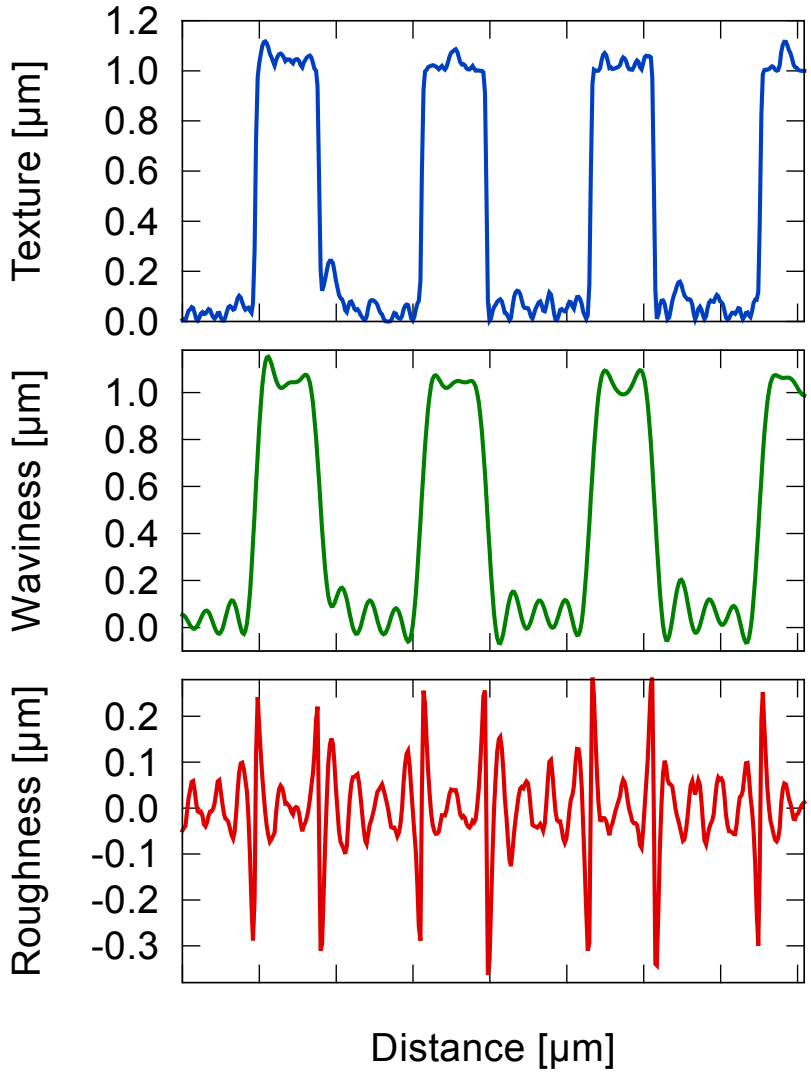


Texture can split to **roughness** and **waviness** (overall shape) using some kind of high-pass and low-pass filters:

- simple splitting in frequency domain
- Gaussian filters
- bidirectional and median filters...



But this is not really means to separate arbitrary overall topography. . .



Irregular areas

Surfaces with different statistical properties in different places represent an even more serious challenge. And this is not covered by norms much.

The first step is **masking** – selecting areas to analyse:

- by an automatic local criterion
- manually
- often a combination of both

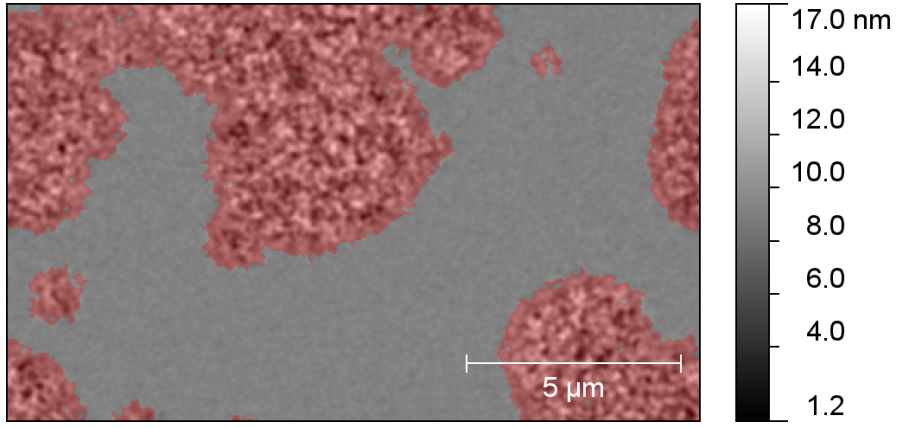
That was the easy part. Now we need to calculate some characteristics.

Simple: quantities based on **height distribution** $\varrho(h)$ alone.

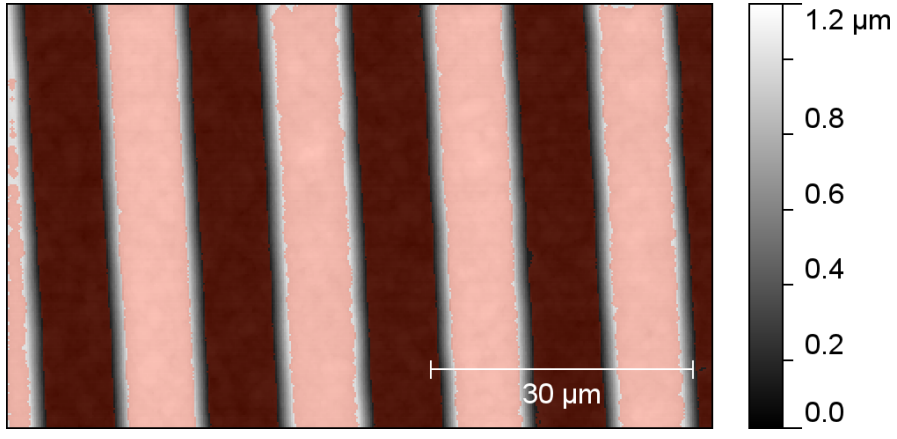
Reasonable: other local quantities (e.g. slope-based).

Difficult: anything else.

But even ACF and PSDF can be calculated – using some math tricks.



	Full	Rough	Smooth
$\sigma =$	1.22 nm	1.85 nm	0.18 nm



	Full	Masked	Grainwise
$\sigma =$	452 nm	490 nm	23 nm

Conclusion

To assess roughness we have to **quantify** it.

Implicit assumption: the surface is ergodic and **the same everywhere**.

Depending on the application, parameters given by a **norm** or the **physical problem** at hand can be used.

The most essential parameters describe the **magnitude** of roughness (S_q, S_a, ...) and are based on **height distribution**.

Spatial parameters are based on **peak counting** (norms) or **correlation properties** (physics).

Root mean square of height deviations and autocorrelation length can be obtained by **fitting** the **experimental ACF** or **PSDF**.

Sampling (discretisation) requires some care to handle properly and define our parameters **sampling-independent**.

If the surface is not the same everywhere we are not completely lost but comparability of results suffers.

Acknowledgements
This work utilised Gwyddion a lot.