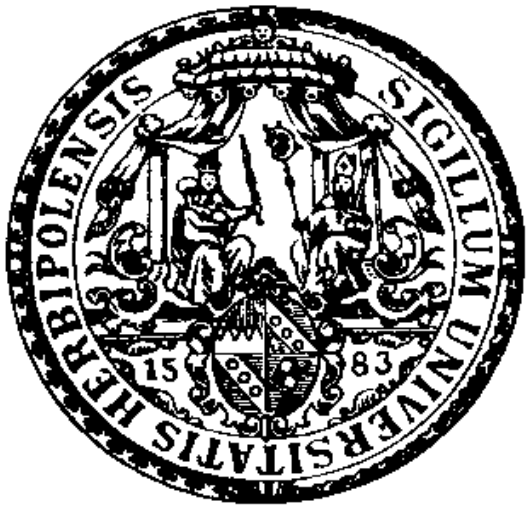


Visual Analysis of Categorical Time Series

Christian H. Weiß

University of Würzburg

Institute of Mathematics

Department of Statistics

This talk is based on the paper

Weiß, C.H.:

Visual Analysis of Categorical Time Series.

Preprint 273, University of Würzburg, 2006.

All references mentioned in this talk correspond to the references in this paper.

Visual Analysis of Categorical Time Series

Motivation

Little existing work deals directly with categorical time series analysis, and much less deals with the visualization of categorical time series. (Ribler, 1997, p. 11)

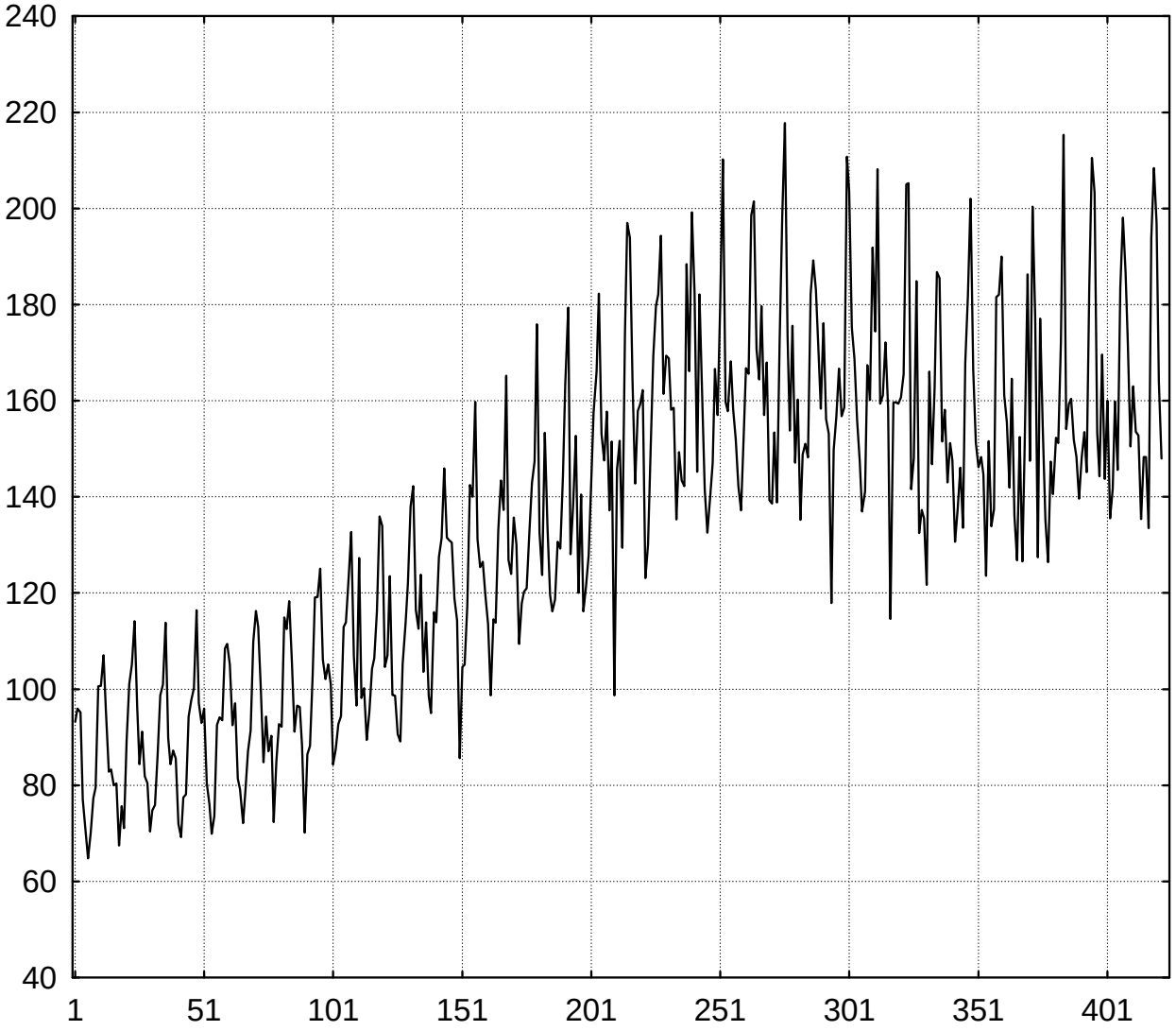
Several visual tools for real-valued time series:

line plot, periodogram, correlogram, etc.

Sometimes criticized: often designed for presentation purposes, and not for exploratory analysis (Unwin, 2000)

Nevertheless: At least line plot simple and universal tool!

Visual Analysis of CaTS – Motivation



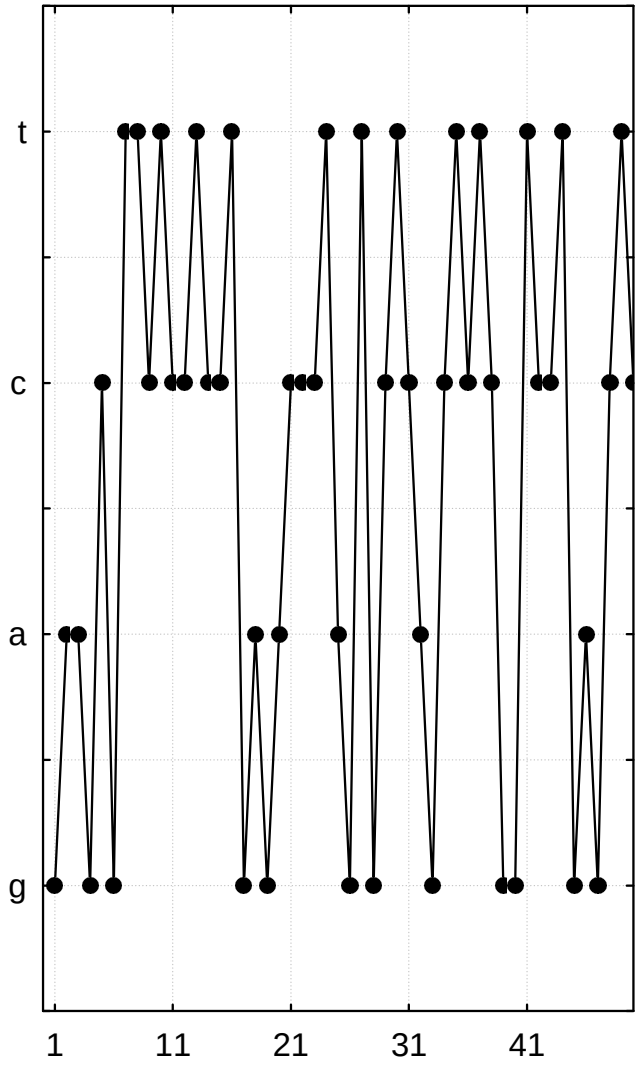
Visual tools for categorical time series: few isolated proposals from computer science and biology.

In fact problematic: Analogue of line plot?

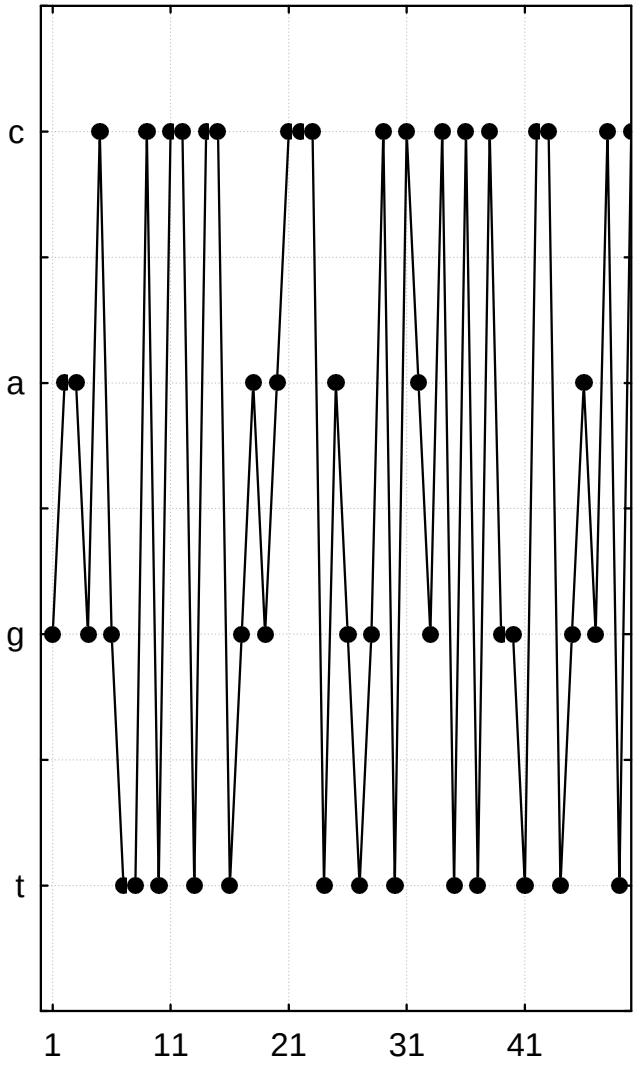
Lack of a natural order within purely categorical range

$\Rightarrow$ arrangement of range along ordinate arbitrary and misleading.

Visual Analysis of CaTS – Motivation



or



Alternative (Keim & Kriegel, 1996): Map range onto set of colors or symbols, plot $x_1, x_2, \ldots$ successively on space-filling curve (line-by-line, column-by-column, Peano-Hilbert curves, spiral, etc.).

However: These graphs perform poorly, characteristic features of time series difficult to recognize, interpretation of resulting plot is problematic, appearance depends heavily on choice of underlying curve.

Genome of Bovine leukemia virus:

So how to analyze categorical time series visually?

Visual Tools for Categorical Time Series

Rate Evolution Graph

Categorical process $(X_t)_{\mathbb{N}}$ with range $\mathcal{V} = \{b_0, \ldots, b_m\}$.

Equivalent representation by binary vectors $\boldsymbol{Y}_t \in \{0,1\}^{m+1}$ with $Y_{t,i} = \delta_{b_i, X_t}$, $i = 0, \ldots, m$.

Define the cumulated sums $\boldsymbol{C}_t := \sum_{s=1}^{t} \boldsymbol{Y}_s$, i. e.,

$$C_{t,i} \; = \text{number of } X_s,\ s = 1, \ldots, t, \text{ equal to } b_i.$$

Rate evolution graph of $(X_t)_{\mathbb{N}}$: (Ribler, 1997)

Multiple line plot of all component series $C_{t,i}$, $i = 0, \ldots, m$, i. e., all $C_{t,i}$ are plotted simultaneously into one chart.

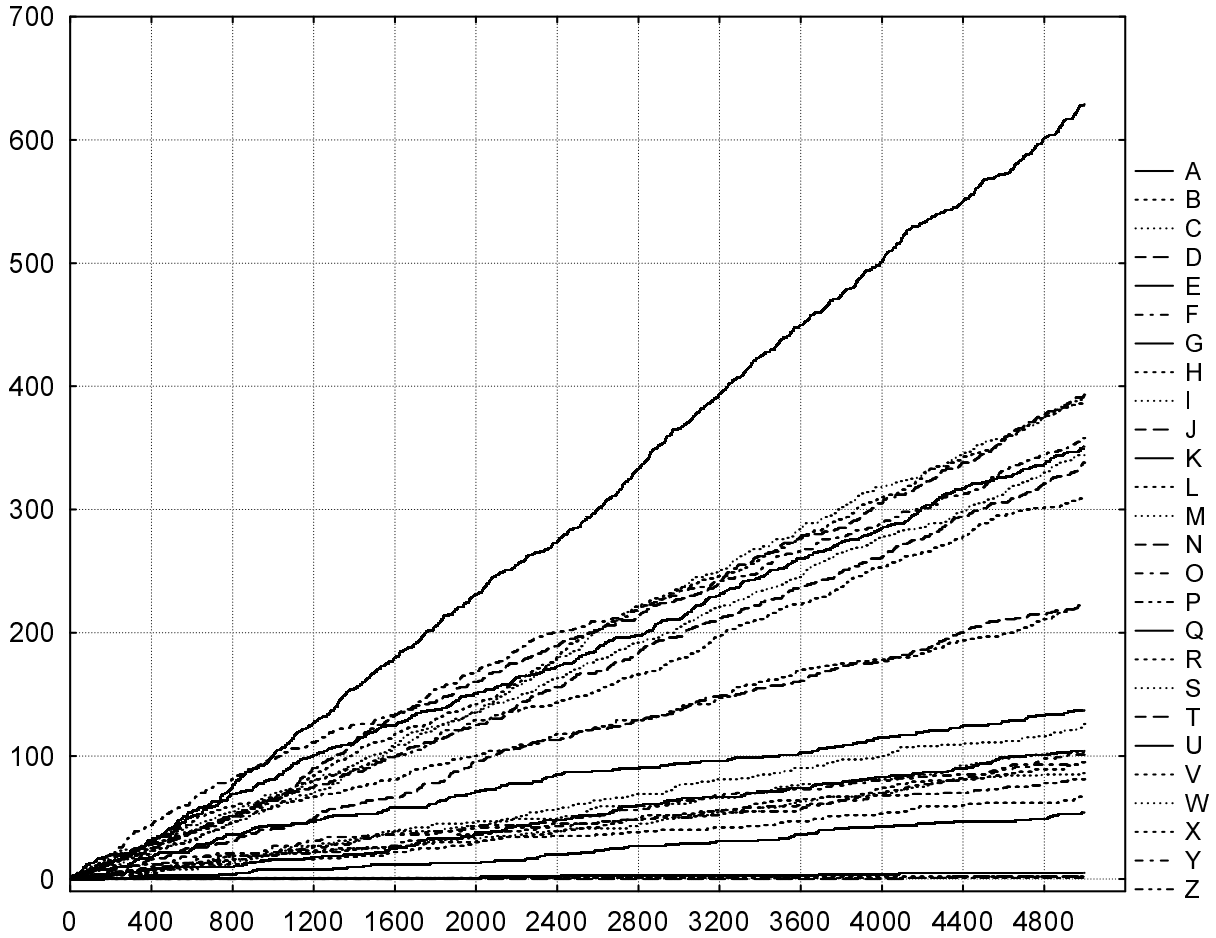
Interpretation:

Slope of graphs is estimate for corresp. marginal probability.

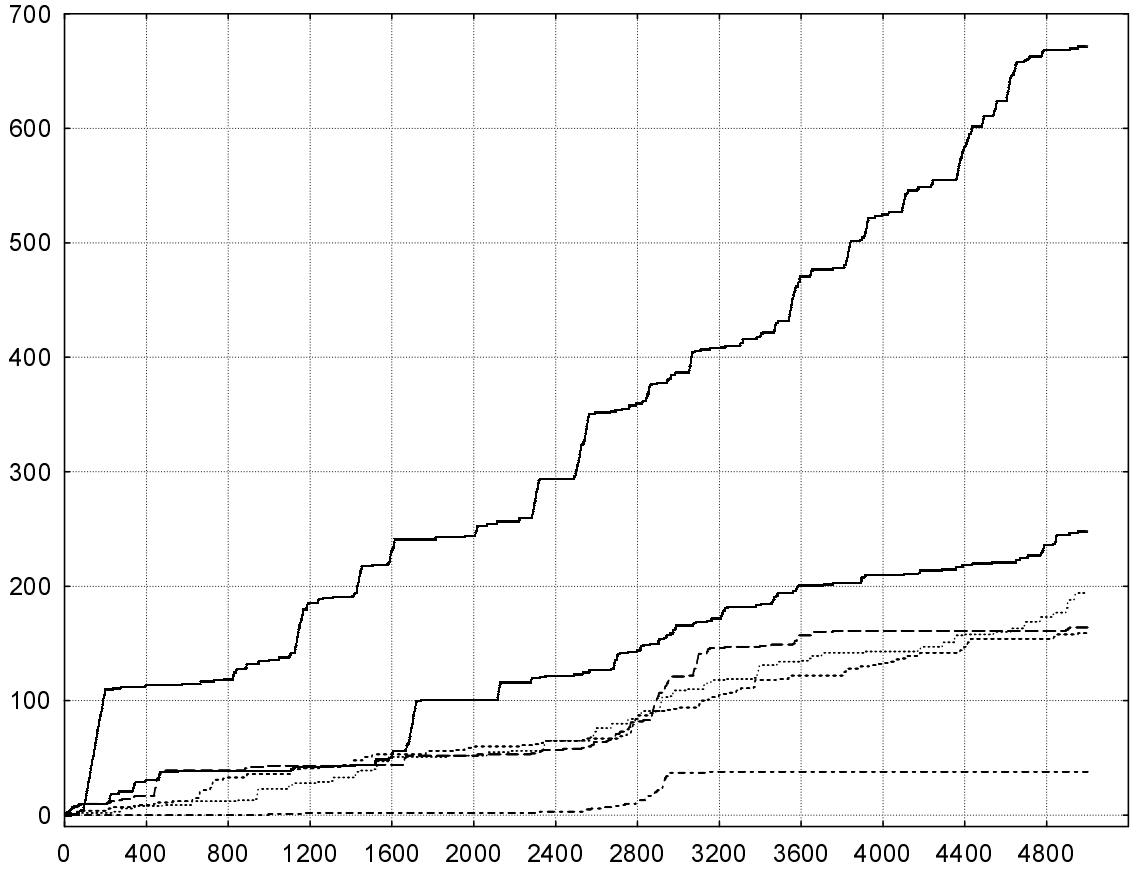
If $(X_t)_{\mathbb{N}}$ marginally stationary and at most moderately serially dependent, then graphs approximately linear in t

$\Rightarrow$ Simple visual tool for checking marginal stationarity.

Shakespeare's (1593) poem "Venus and Adonis":



Log data of Statistics server: Access to home directory of five members.



Visual Tools for Categorical Time Series

IFS Circle Transformation

Important task of categorical time series analysis ('sequential pattern analysis'):

Frequency of tuples of symbols (strings).

Possible approach:

IFS Circle Transformation of Weiß & Göb (2005).

Categorical process $(X_t)_{\mathbb{N}}$ with range $\mathcal{V} = \{b_0, \ldots, b_m\}$.

IFS Circle Transformation consists of two steps:

1. Recode range by real vectors (**circle transformation**):

$$\boldsymbol{b}(b_k) = (\cos(k \cdot \tfrac{2\pi}{m+1}), \sin(k \cdot \tfrac{2\pi}{m+1}))^\top, \qquad \boldsymbol{Z}_t := \boldsymbol{b}(X_t).$$

2. Generate **fractal series** $\boldsymbol{Y}_0 = \boldsymbol{y}_0, \boldsymbol{Y}_1, \boldsymbol{Y}_2 \ldots$ (online!):

$$\boldsymbol{Y}_t \;=\; \alpha \boldsymbol{Y}_{t-1} \;+\; \beta \boldsymbol{Z}_t, \quad \text{where } \boldsymbol{y}_0 = 0,$$

with weights $0 < \alpha < 1$, $\beta > 0$.

Observation:

If parameter α is chosen appropriately, then:

- If two segments $(X_i, ..., X_{i-k}) = (X_j, ..., X_{j-k})$, then points $\boldsymbol{Y}_i$ and $\boldsymbol{Y}_j$ will be *close* in $\mathbb{R}^2$,

- but if the segments differ, then points $\boldsymbol{Y}_i$ and $\boldsymbol{Y}_j$ will be *distant* in $\mathbb{R}^2$.

$\Rightarrow$ Screen fractal series $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_T, ...$ for *close* elements (cluster). Points in a cluster correspond to occurrences of similar patterns in $X_1, ..., X_T$.

Theorem of Weiß & Göb (2005):

If $\alpha < \frac{d_{m+1}}{2(d_{m+1}+3)}$, $d_n = \sqrt{2(1-\cos\frac{2\pi}{n})}$, then equivalence:

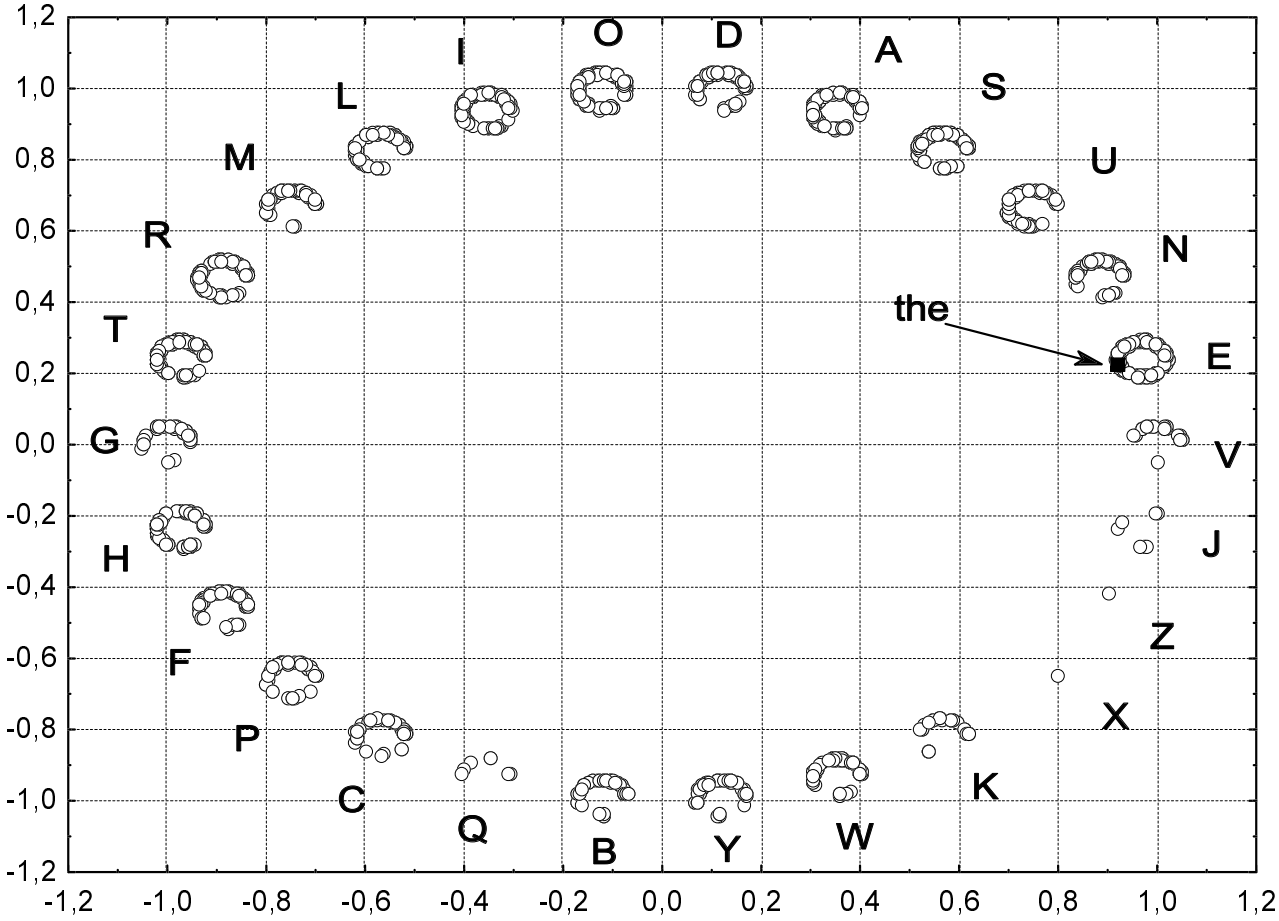
$$\|\boldsymbol{Y}_i - \boldsymbol{Y}_j\| < d_{m+1}\,\frac{\beta}{3}\,\alpha^k$$

$$\Longleftrightarrow \quad (X_i,\ldots,X_{i-k}) = (X_j,\ldots,X_{j-k}).$$

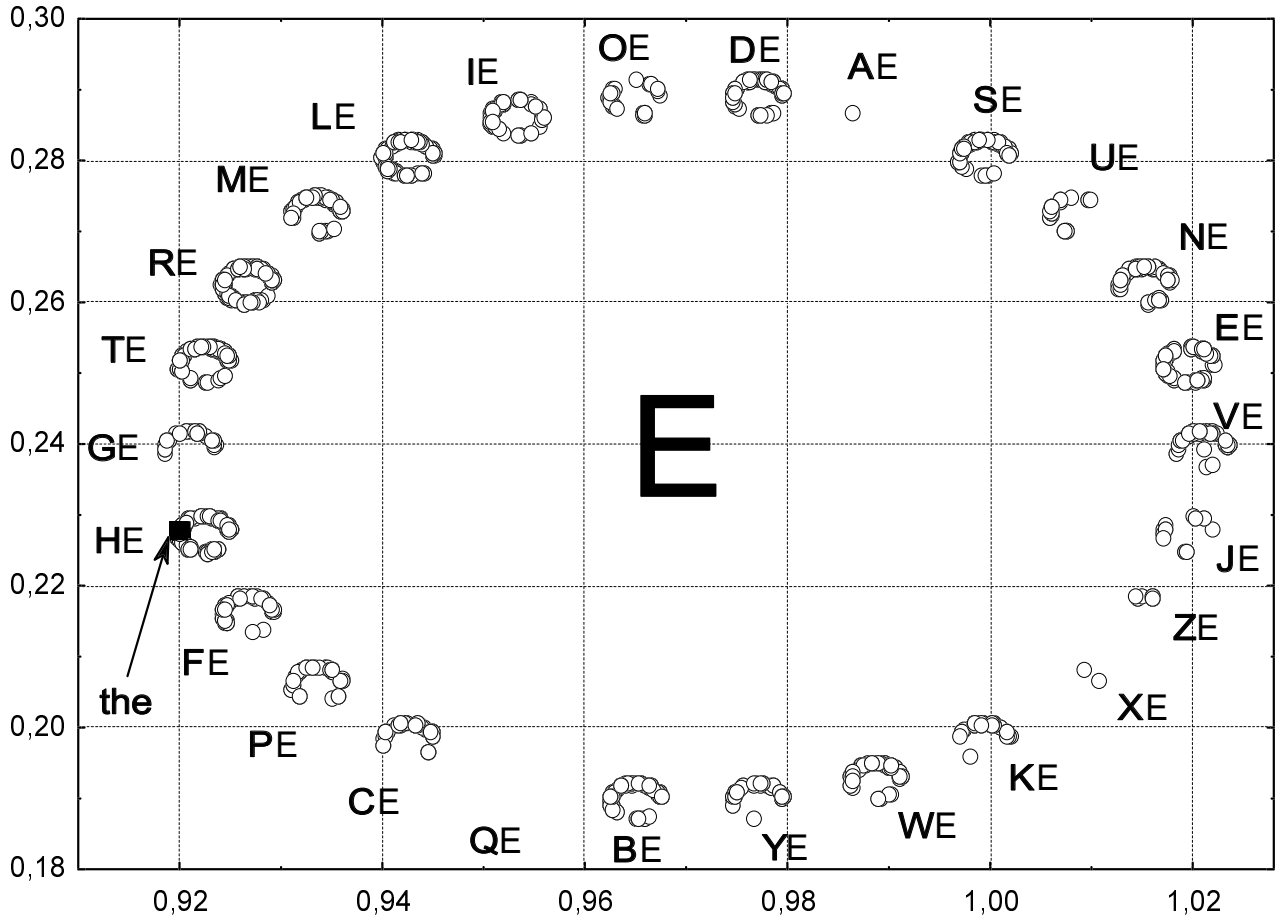
In contrast, if $(X_i,\ldots,X_{i-(k-1)}) = (X_j,\ldots,X_{j-(k-1)})$ but $X_{i-k} \neq X_{j-k}$, then

$$\|\boldsymbol{Y}_i - \boldsymbol{Y}_j\| \geq d_{m+1}\,\frac{\beta}{3}\,\alpha^k.$$

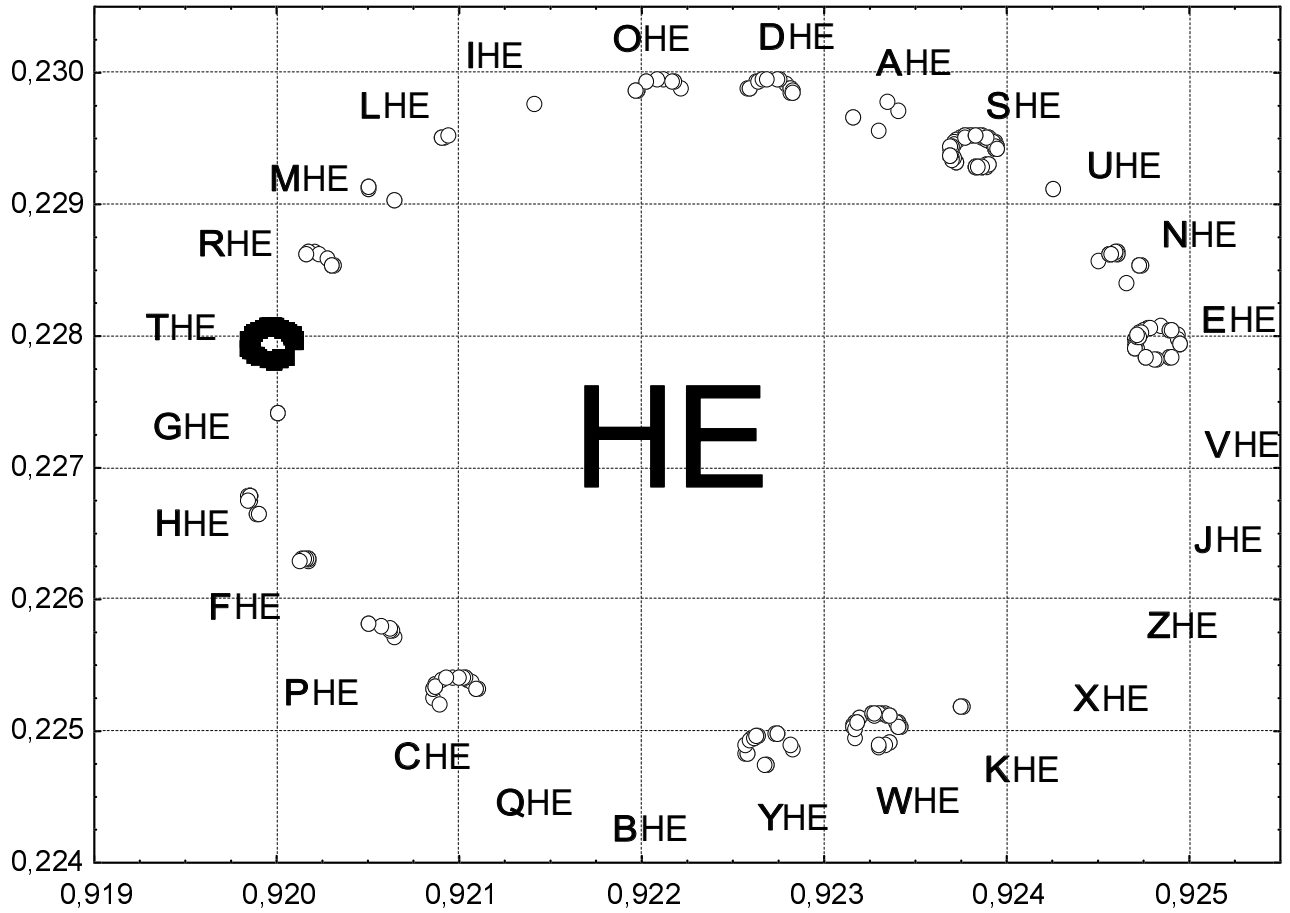
Shakespeare's (1593) poem "Venus and Adonis":



Shakespeare data around last letter ‘E’:



Shakespeare data around last letters ‘HE’:



Visual Tools for Categorical Time Series

Pattern Histograms

Idea: Analyze frequency distribution of a certain categorical feature (runs, cycles).

A **cycle**, beginning at time t with $X_t = x_t$, is closed at time $t + k > t$ iff $X_t = x_t = X_{t+k}$ but $X_s \neq x_t$ for all $t < s < t + k$.

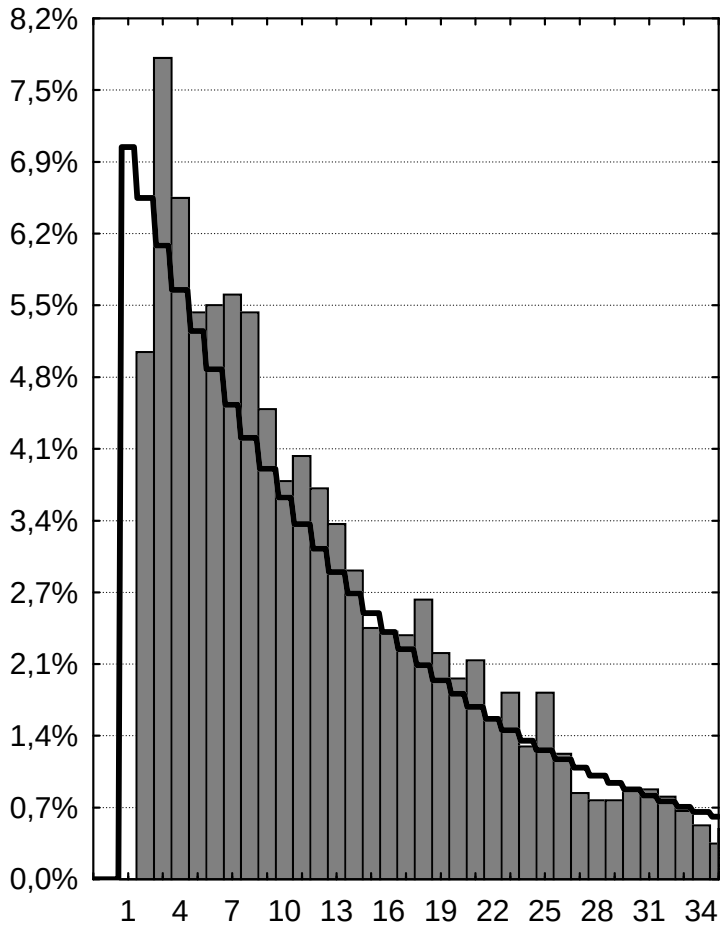
By definition, a new cycle starts at any time t, and a previous cycle is closed.
The **length of a cycle** starting in t and ending in $t + k$ is defined as k.

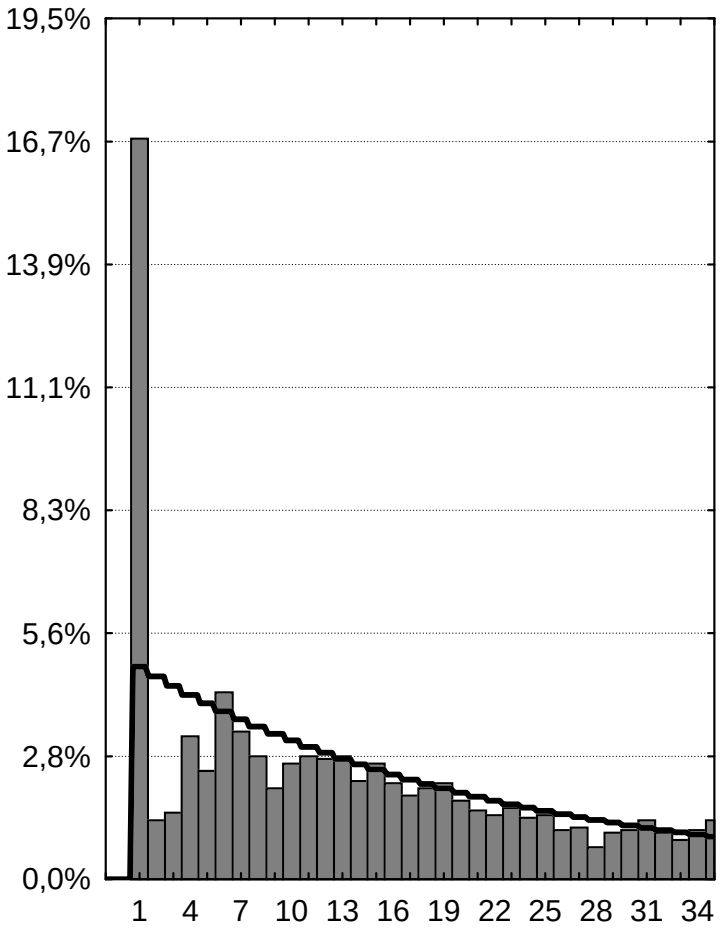
Procedure: Given $(x_t)_{t=1,2,\ldots}$. For each $b_i \in \mathcal{V}$, count runs/cycles, grouped according to their length. Then construct histogram for each $b_i \in \mathcal{V}$.

$\Rightarrow$ **Pattern histograms** estimate underlying run/cycle length distribution, which depends on both concrete symbol b_i and true process model $\Rightarrow$ **Pattern histograms** help to identify the model or to check model hypothesis.

Example: If $(X_t)_{\mathbb{N}}$ stationary Bernoulli process, then run/cycle lengths related to geometric distribution.

‘a’- and ‘l’-Cycles in Shakespeare’s “Venus and Adonis”:





Visual Tools for Categorical Time Series

Categorical Control Charts

Statistical process control aims at monitoring and improving production processes. Most popular tool: **Control chart**, helps to identify a deviation from the state of control.

Control charts as tool of exploratory data analysis: **Control charting in phase I** is retrospective analysis of historical process data. It helps to check the stationarity and model class assumption, and to identify an appropriate process model.

Categorical Control Charts monitor certain categorical feature, e. g.:

- Run or cycle lengths of certain symbol $b_i \in \mathcal{V}$,
- marginal process distribution, etc.

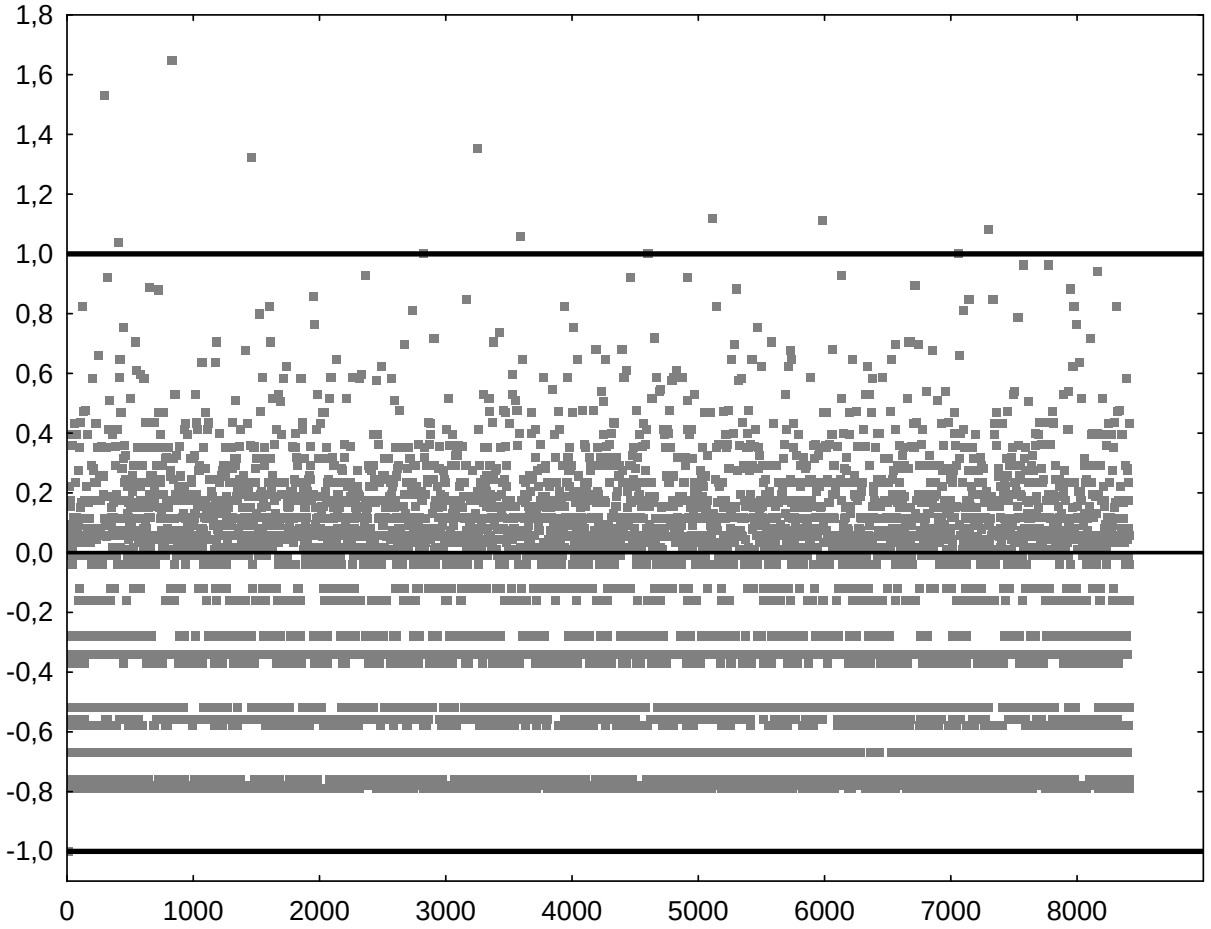
Design of corresponding chart (center line, control limits) is based on hypothetical process model.

Due to time restrictions:

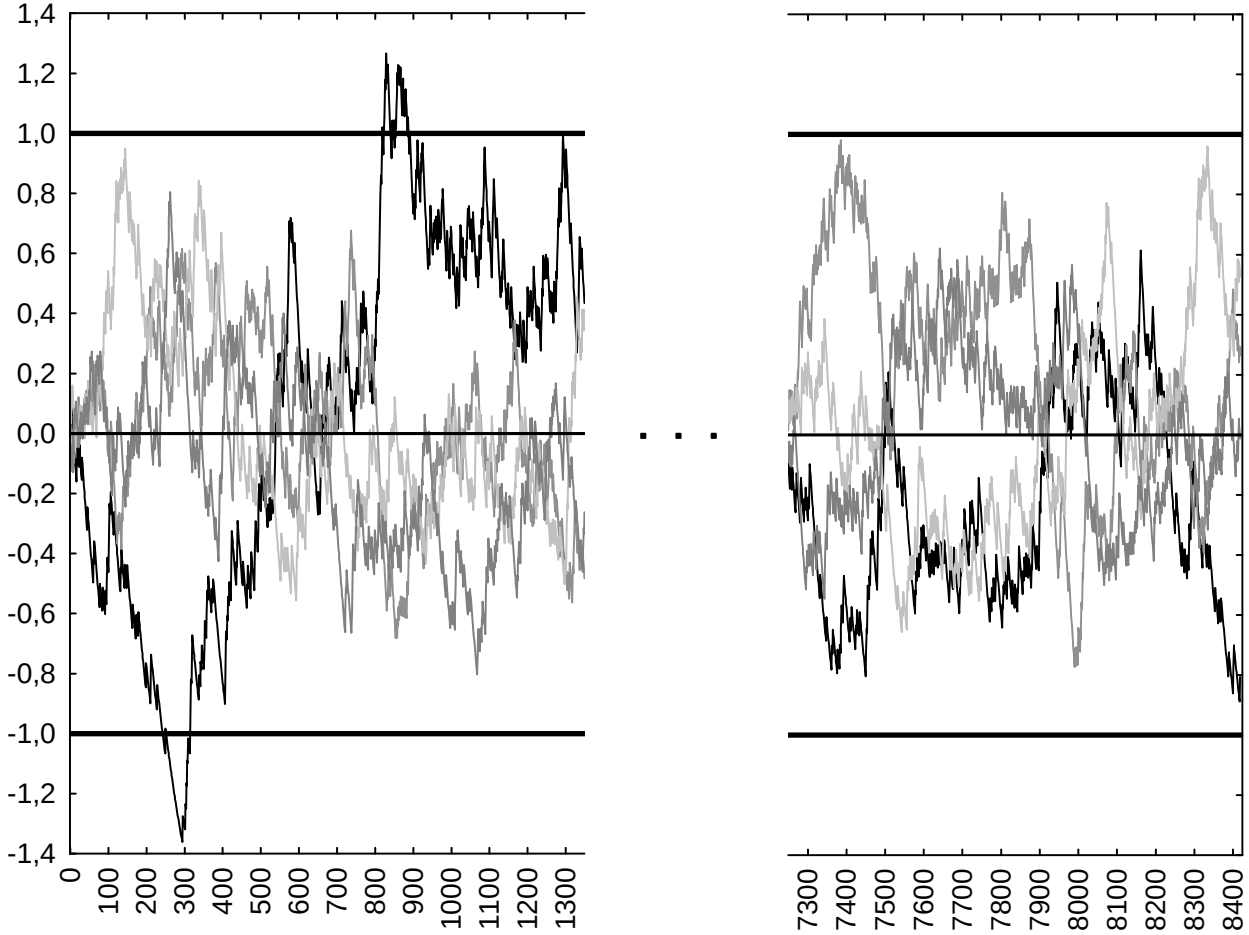
Details and examples in article of Weiß (2006).

Impressions of possible charts on the following slides . . .

Control chart of cycle lengths in Bovine data:



EWMA control chart of Bovine data:



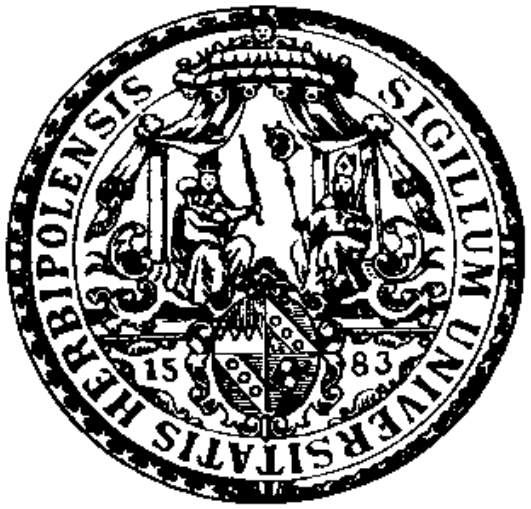
Visual Tools for Categorical Time Series

Summary

Presentation of known and new visual tools, especially:

- rate evolution graph to check stationarity assumption,
- IFS circle transformation for sequential pattern analysis,
- pattern histograms for model identification,
- control charts for stat. analysis and model identification.

However: These tools are very specialized.
Universal instrument (like line plot), providing multiple types of information at once, still missing.



Thank You

for Your Interest!

Christian H. Weiß

University of Würzburg

Institute of Mathematics

Department of Statistics