Geographic Information Technology Training Alliance (GITTA) presents:

Thematic Change Analysis

Responsible persons: Chloé Barboux, Claude Collet

Table Of Content

1. Thematic Change Analysis

How can thematic changes of spatial features be analysed?

According to the wide diversity of contexts presented there are numerous potential methods for this task. Objectives of such analysis concentrate on the change of thematic properties of observations – in our case spatial features - regardless of their spatial nature and distribution. Of course these methods can pretend to belong to the toolbox of spatial analysts. However they need companion methods applied in a further stage of analysis to fulfil the task of spatial analysis: the spatial distribution analysis of these thematic changes. These companion methods will be presented in the Unit 3: Spatial change analysis.

The following criteria defining the context of analysis are considered:

- *Time description*: 2 limits or discrete intervals
- *Level of measurement*: qualitative or quantitative
- *Single or multiple* observations analysis
- *Uni or Multi-variate* analysis

The presentation of thematic change analysis methods is organised into 3 groups of methods:

-**Production of change indices:** univariate description of a set of multiple observations

-**Time series behaviour description:** univariate description for individual behaviour of observations

-**Multivariate methods:** multivariate description of a set of multiple observations.

Let's take a common example to illustrate thematic change analysis discussed in this Lesson. The study area is composed of 9 spatial features corresponding to the municipalities of a district. Several phenomena were measured during a period of time ranging from 1900 to 1990.

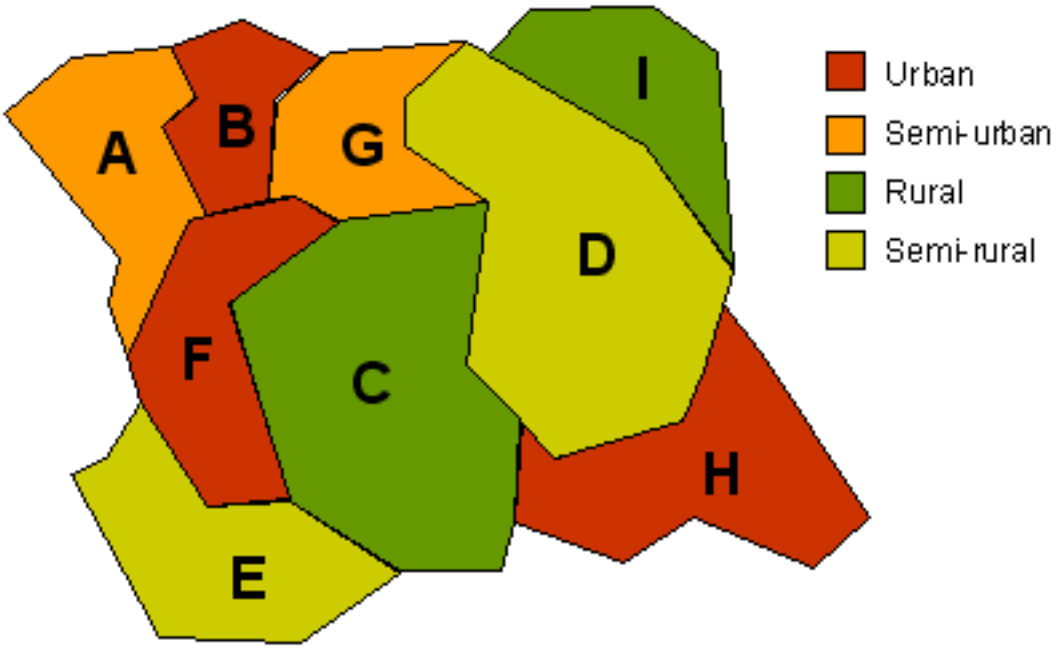
Variables expressing their properties at each decade during this period of time are the following:

-Number of inhabitant (population): quantitative data at cardinal level

-Political majority: qualitative data at nominal level

-Number of wired telephone subscription: quantitative data at cardinal level

-Gender of the municipality mayor: qualitative data (binary) at nominal level



The 9 municipalities constituting the district under study

Municipality	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990
A	615	701	757	821	921	969	969	1156	1500	2118
B	3453	3956	4160	3259	3634	4033	4248	4932	5568	6298
C	523	561	546	600	576	563	448	532	573	612
D	683	661	650	690	768	770	812	1084	1180	1322
E	311	328	339	354	363	488	761	1303	1121	1700
F	774	899	660	822	1066	1334	1813	4329	5245	6469
G	964	1280	1267	1436	1463	1570	1621	2021	2255	2316
H	856	907	1021	1234	1317	1490	2584	5214	5753	7883
I	87	95	81	83	95	81	52	58	45	59
District	8266	9388	9481	9299	10203	11298	13308	20629	23240	28777

Number of inhabitant at each decade during the period 1900 to 1990

Learning Objectives

- You will be able to select appropriate methods and techniques for a specific context of thematic change analysis
- You will master the basic principles of various thematic change methods and be able to access to related text references for in depth learning

1.1. Production of change indices

How to summarise changes of thematic properties?

There are many ways to describe change through indices. With this approach property change of individual spatial features is summarised with an index. This can account for an absolute or relative change, in an univariate or multivariate context. The following table presents various methods for the production of *change indices* [1].

	TIME DIMENSION	
	2 limits	**Intervals**
UNIVARIATE	**Production of change indices:**	
	Property differencing (Quant)	Central tendancy / variability (Qual/Quant)
Multiple Observations	Property ratioing (Quant)	Regression score (Quant)
	Cross-tabulation / classification (Qual)	Allometric score (Quant)
	Transition matrices (Qual)	Standardised PCA scores (Quant)

Qual: qualitative data (nominal) Quant: quantitative data (ordinal, cardinal)

A brief overview of change indices methods

Let us first differentiate between methods dealing with only 2 limits of time and those managing several discrete intervals.

1.1.1. Methods processing 2 time markers (2 limits)

In many situations the time dimension is only describe through the status of observation properties at the beginning and the end of the considered time period (t_{min}, t_{max}). Assuming a context of a univariate description of multiple observations, information can be structured as a two dimensional table with rows corresponding to observations and columns expressing the two time limits.

[1] A change index is an indicator derived from multitemporal measurements. It expresses the amount of change within a period of time. It can describe the change behavior of a set of features (global) or of individual features. It can result from a difference, a ratio, ...

Number of inhabitant in 1900 and 1990

Municipality	1900	1990
A	615	2118
B	3453	6298
C	523	612
D	683	1322
E	311	1700
F	774	6469
G	964	2316
H	856	7883
I	87	59
District	8266	28777

Political majority in 1900 and 1990

Municipality	1900	1990
A	2	1
B	4	4
C	2	2
D	2	2
E	1	4
F	1	3
G	4	4
H	3	4
I	2	1

Liberal:1, Republican:2, Democratic:3, Socialist:4

Number of inhabitant and political majority in 1900 and 1990

As stated previously, property change can be considered either individually or globally among all spatial features or individually but relatively to the global behaviour of features.

Individual property change (comparison of the two properties):
As stated previously, property change can be considered either individually or globally among all spatial features or individually but relatively to the global behaviour of features.

Property difference (quantitative)
D index is the difference between property value V at t_{max} and t_{min} for each i spatial feature:

$$D_i = V_{i,\, t_{max}} - V_{i,\, t_{min}}$$

The normalised difference ND expresses the change rate based on a reference time (often t_{min}) for each i spatial feature:

$$ND_i = (V_{i,\, t_{max}} - V_{i,\, t_{min}}) / V_{i,\, t_{min}}$$

Property ratio (quantitative)
R index is the ratio between property value V at t_{max} and t_{min} for each i spatial feature:

$$R_i = V_{i,\, t_{max}} / V_{i,\, t_{min}}$$

Another property ratio can be produced that allows comparison between individual change rate and the global one (for the whole set of considered features), R_{tot}. This relative change rate index RR_i (Relative Ratio) is the ratio between R_i and R_{tot}:

$$RR_i = R_i / R_{tot}$$

Change classification (qualitative)

C index expresses the type of change between property value V at t_{max} and t_{min} for each i spatial feature. C values can be simply binary (with 0 for no change and 1 for change) or multiple to describe the type of change between the two considered categories. C results from a classification process.

In the case of nominal level content, C value is either 0 or 1 expressing a change or not. However, for variables at ordinal or cardinal level, one might want to differentiate between three situations of change: a decrease, no change and an increase. The 3 possible values of C index could then be derived from the classification (recoding) of D_i values, according to the following scheme:

- IF $D_i < 0$ then $C_i = 1$
- IF $D_i = 0$ then $C_i = 2$
- IF $D_i > 0$ then $C_i = 3$

Illustration

Let us apply this set of individual property change indices to the two variable sets listed in **Table** above. The change of individual features between the two intervals of time is described as follow.

For the number of inhabitant (quantitative)

Municipality	1900	1990	D(90,00)	ND(90,00)	R(90,00)	RR(90,00)	C(90,00)
A	615	2118	1503	0.710	3.444	0.989	3
B	3453	6298	2845	0.452	1.824	0.524	3
C	523	612	89	0.145	1.170	0.336	3
D	683	1322	639	0.483	1.936	0.556	3
E	311	1700	1389	0.817	5.466	1.570	3
F	774	6469	5695	0.880	8.358	2.401	3
G	964	2316	1352	0.584	2.402	0.690	3
H	856	7883	7027	0.891	9.209	2.645	3
I	87	59	-28	-0.475	0.678	0.195	1
District	*8266*	*28777*	*20511*	*0.713*	*3.481*	*1.000*	*3*

For the political majority (qualitative)

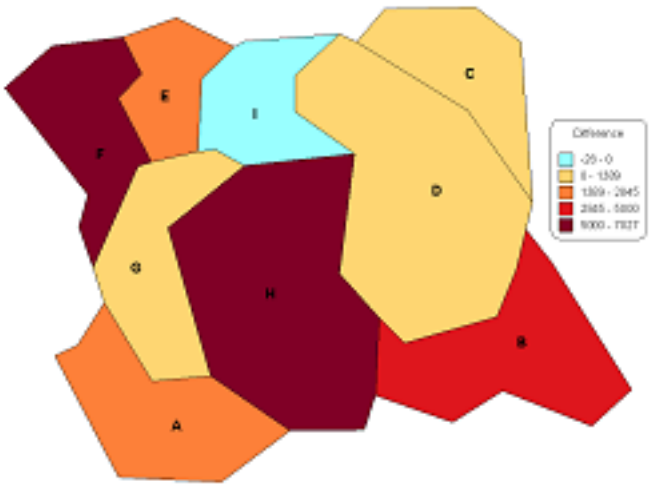
Municipality	1900	1990	C(90,00)
A	2	1	1
B	4	4	0
C	2	2	0
D	2	2	0
E	1	4	1
F	1	3	1
G	4	4	0
H	3	4	1
I	2	1	1

0: no change, 1: change

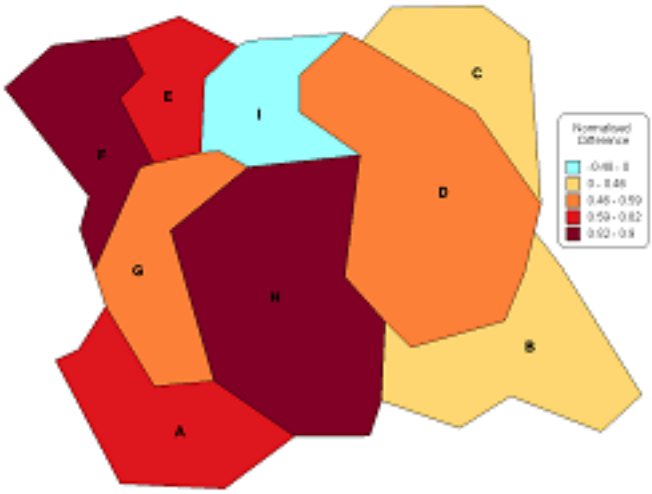
In the example it shows that R_i amplifies growth changes compared to ND_i, but the reference value for a decrease is now less than 1 (municipality I).
The spatial analysis of behaviour change can then be carried on as a further step by making use of index mapping.

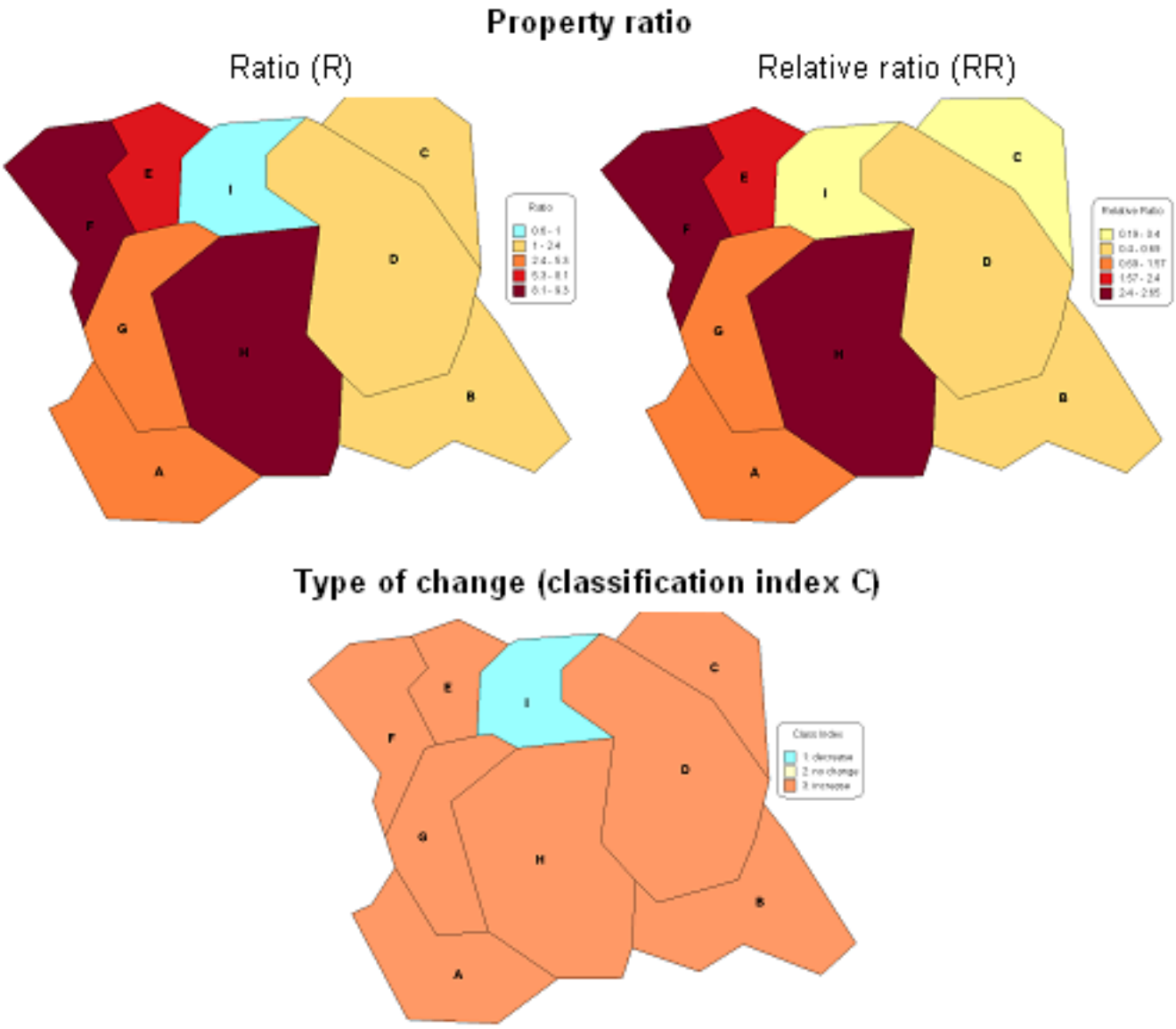
Property difference

Difference (D)

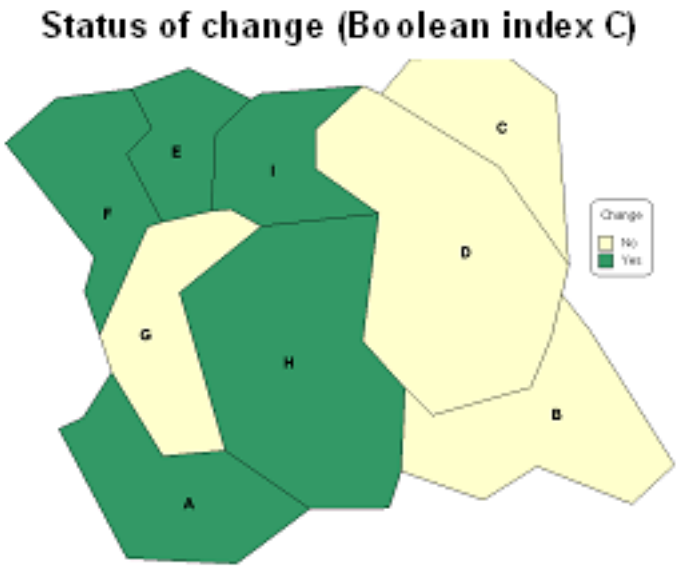


Normalised difference (ND)





Mapping of property change indices for the number of inhabitant (quantitative)



Mapping of property change index for the political majority (qualitative)

Global property change:
One can be interested in evaluating changes among the whole set of spatial features.

Summary statistics (quantitative / qualitative)
Individual change indices can be summarised using appropriate central tendency and dispersion indicators. Of course only change indices relevant for comparison between features should be considered.
For quantitative change indices (ordinal and cardinal levels):

- Mean or median indicators, applied to the normalised difference index (ND) for example
- Standard deviation or interquartile of the normalised difference index (ND) values for example
- Median or mode indicators, applied to the change classification index (C) for example
- Interquartile or diversity indicators, applied to the change classification index (C) values for example

For qualitative change indices (nominal level):

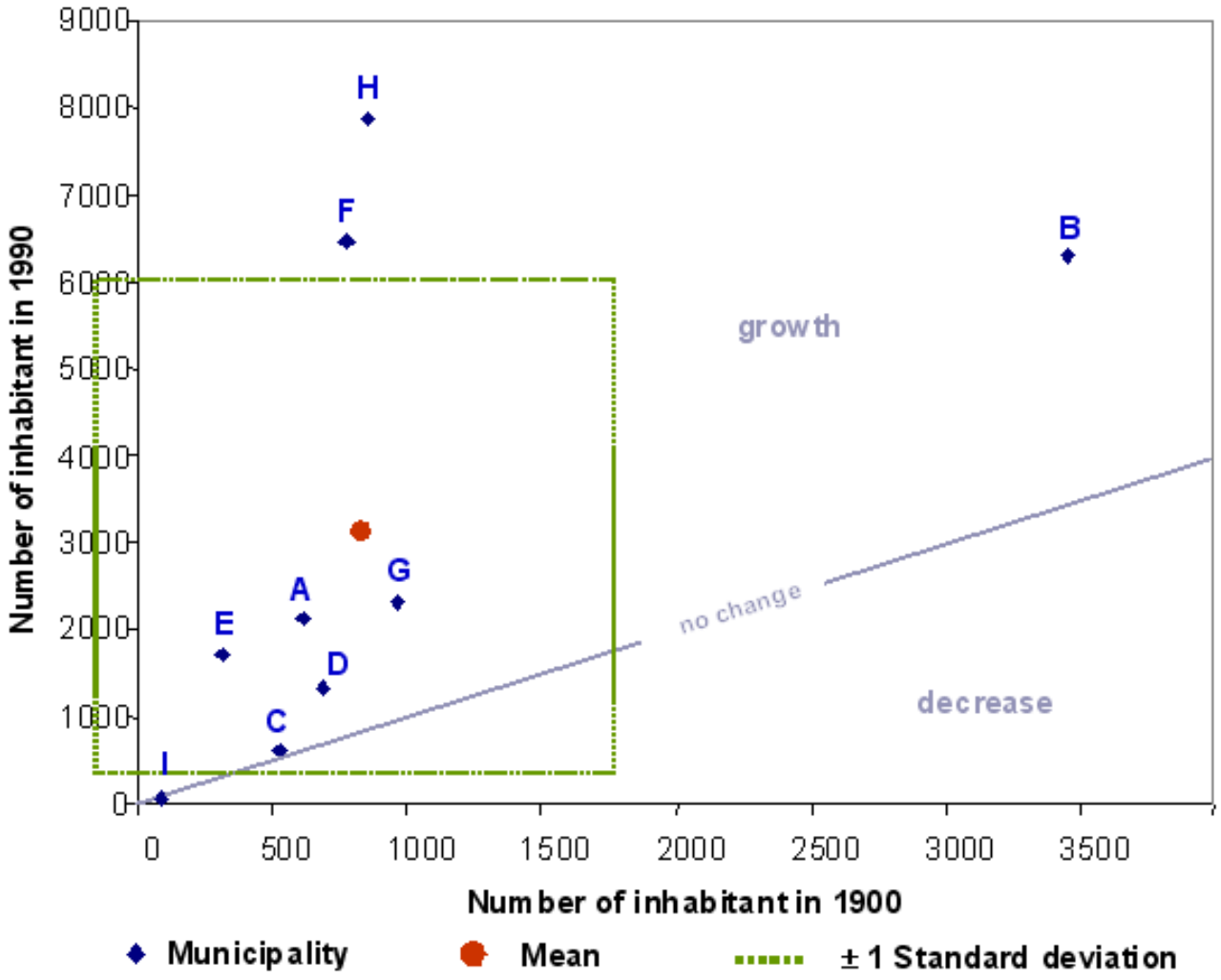
- The mode, applied to the change classification index (C) for example
- The diversity index of the change classification index (C) values for example

These relevant central tendency and dispersion descriptors can then be used for a relative description of individual change behaviour. At ordinal and cardinal levels individual feature change can be compared to the global change by grouping its change value into classes around the central tendency:
- at ordinal level: interquartiles around the median;
- at cardinal level: standard deviation units around the mean value.

Scattergramme (quantitative)
Another efficient way for comparison of change between features is to plot them according to their individual properties for the two dates. This can be seen as a "*time change map*" with a pair of coordinates locating them into this two-dimensional space.
Such a diagramme allows different types of interpretation:

- comparison between observations: the distance expresses the level of similarity
- when adding the location of the central tendency (mean, median) and the variability (standard distance) to the diagramme, comparison can be performed between each observation and these references
- furthermore change behaviour references can be added to the graph in order to identify three types of change: growth, no change and decrease. These types of change correspond to three areas on the diagramme: on the diagonal, above and below.

Next figure illustrates the use of a diagramme representation for the mapping of the 9 municipalities. Their change behavior can be compared to each other as well as to global references (mean and standard deviation) and types of change.



Graphical representation of property change for the number of inhabitant using a scattergramme as a "time change map"

EXERCISE

On the last figure:

- identify groups of municipalities with a similar change behaviour
- identify municipalities away from the average behaviour. How do you interpret their position?
- Which municipality belongs to the "decrease" type of change?

- Compare the distance of each municipality position to the "no change" line. To what individual property change index listed in tables of the Illustration paragraph can it be best associated?

Transition matrices (qualitative /quantitative)

Now our interest is in the nature of transitions from one state to another. We can use techniques that sacrifice all information about individual observation properties but provide in return information on the tendency of one state to follow another.

Due to the use of a two dimensional cross-table, the number of considered properties is limited. This approach is fully adapted for qualitative data (categories at nominal level), but also for classes at ordinal level and even at cardinal level, assuming original properties are grouped into a limited number of classes.

Let us illustrate the exploitation of *transition matrices* [2] with our data set on Political majority of municipalities in 1900 and 1990 illustrated in previous table. We would like to summarise the change from the property in 1900 to another one in 1990. Furthermore we could identify the tendency of one property to follow another.

A 4 × 4 matrix (or cross-table) can be constructed showing the number of times a given property –political majority- is succeeded by another, a matrix of this type is called a ***transition frequency matrix*** and is shown in the following table. In order to avoid confusion between properties and frequencies of change patterns, let us recode property values with letters (L for Liberal, R for Republican, D for Democratic and S for Socialist). The considered sample contains 9 observations, so there are 9 transitions. The matrix is read from rows to columns meaning, for example, that a transition from state $_{L}$ to state $_{D}$ is counted as an entry element $a_{1,3}$ of the matrix. That is if we read from the row labelled $_{L}$ to the columns labelled $_{D}$, we see that we move from state $_{L}$ into state $_{D}$ one time in the set of observations, but we can observe that there is no occurrence of move from state $_{D}$ into state $_{L}$ (entry element $a_{3,1}$). The transition frequency matrix is asymmetric and in general $a_{i,j}$ # $a_{j,i}$. The transition frequency matrix is a concise way of expressing the incidence of one state or property following another, the *transition pairs*.

		To				*Row totals*	
		L	**R**	**D**	**S**		
	L	0	0	1	1	*2*	
From	**R**	2	2	0	0	*4*	
	D	0	0	0	1	*1*	
	S	0	0	0	2	*2*	
Column totals		*2*	*2*	*1*	*4*	**9**	***Grand total***

A transition frequency matrix showing property change patterns in political majority between 1900 and 1990

The tendency for one state to succeed another can be emphasised in the matrix by converting the frequencies to decimal fractions or percentages. Different types of relative frequencies can be derived:

[2] A general term to identify any matrix that expresses a change of properties (states) between two moments

- If each element is divided by the Grand total, the resulting fractions express its relative frequency of occurrence. The whole matrix then shows the relative frequency of all the possible types of transitions. Such a matrix is called ***transition relative frequency matrix*** .
- If each element in the ith row is divided by the total of the ith row, the resulting fractions express the relative number of times state i is succeeded by the other states. Such a matrix is called *transition proportion matrix* [3]. The advantage of this matrix is to express the tendency of one state to follow another regardless of the total occurrence of the initial state.

		To				Row totals	
		L	R	D	S		
	L	0	0	0.11	0.11	*0.22*	
From	R	0.22	0.22	0	0	*0.44*	
	D	0	0	0	0.11	*0.11*	
	S	0	0	0	0.22	*0.22*	
Column totals		*0.22*	*0.22*	*0.11*	*0.44*	***≅ 1.0***	***Grand total***

The transition relative frequency matrix showing all the possible types of transitions in political majority between 1900 and 1990

		To				Row totals
		L	R	D	S	
	L	0	0	0.5	0.5	*1.0*
From	R	0.5	0.5	0	0	*1.0*
	D	0	0	0	1.0	*1.0*
	S	0	0	0	1.0	*1.0*

The transition proportion matrix showing the tendency of one state of political majority in 1900 to follow another in 1990

The transition relative frequency matrix shows that 44% of the municipalities had the property R in 1900 and this proportion has decreased to 22% in 1990. This 22% of loss compensates for the loss of state L during the same period. The property S has the highest proportion of resulting states with 44% of municipalities. On the other hand the indicates the same tendency for state L to move to states D and S (L/D and L/S = 0.5), as opposed to the state R where the tendency of unchanged is 0.5 (R/R).

Assuming a representative sample of features, relative frequencies can be interpreted as probabilities of occurrence. This extended approach can make use of Markov chains for estimating the probability of occurrence of a state based on the existence of a previous stage. This method will be discussed in the next section relative to time series analysis of a sequence of data. It can be used to describe individual transition pattern.

[3] A matrix that expresses the tendency of one state to follow another

EXERCISE

Compare the two last tables and apply the same reasoning for the final state S.

1.1.2. Methods for time series

Evolution of thematic properties throughout time can be described in greater details by measuring the property of individual observation (each spatial feature) at numerous intervals of time during the considered period. Time intervals can be either regular or irregular. However many methods require a same regular interval that is common to all observations in order to allow comparison. For each observation this sequence of property values is called a *time series* [4]. Time series collected for a set of observations can be structured as a two-dimensional data table as illustrated below.

Municipality	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990
A	615	701	757	821	921	969	969	1156	1500	2118
B	3453	3956	4160	3259	3634	4033	4248	4932	5568	6298
C	523	561	546	600	576	563	448	532	573	612
D	683	661	650	690	768	770	812	1084	1180	1322
E	311	328	339	354	363	488	761	1303	1121	1700
F	774	899	660	822	1066	1334	1813	4329	5245	6469
G	964	1280	1267	1436	1463	1570	1621	2021	2255	2316
H	856	907	1021	1234	1317	1490	2584	5214	5753	7883
I	87	95	81	83	95	81	52	58	45	59
District	8266	9388	9481	9299	10203	11298	13308	20629	23240	28777

Change in number of inhabitant during the period 1900 – 1990 with an interval of 10 years

Municipality	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990
A	2	1	1	2	2	4	3	3	3	1
B	4	4	4	4	2	2	4	4	4	4
C	2	3	3	2	1	2	1	1	2	2
D	2	2	2	1	1	1	2	2	1	2
E	1	1	2	1	2	2	3	3	4	4
F	1	1	3	3	3	3	4	4	3	3
G	4	4	4	2	2	3	3	4	4	4
H	3	3	3	4	4	3	4	3	3	4
I	2	2	1	2	1	1	1	1	1	1

Liberal:1, Republican:2, Democratic:3, Socialist:4

Change in political majority during the period 1900 – 1990 with an interval of 10 years

[4] A sequence of measurements ordered according to Time (moments of time). It describes the change of properties of a single observation throughout time

Individual property change:
The objective is again to summarise the individual time change behaviour with a single index value. This can be done by the use of typical statistical descriptors (central tendency and dispersion) or by modelling the trend in change.

Summary statistics
With a simple statistical descriptor some aspects of the individual change behaviour can be described such as its central tendency, its variability or a combination of both. Depending on the nature of data analysed (qualitative or quantitative), the following statistical descriptors are applied to summarise the change:

- *For qualitative data time series:* the mode and the diversity
- *For quantitative data time series:* the mean or the median, the standard deviation or the interquartile, the coefficient of variation and the range.

Let us illustrate the use of statistical descriptors for summarising the time change behaviour of features described in the two tables above.
The time series describing the political majority during the period 1900-1990 for 9 municipalities are qualitative data measured at nominal level, therefore two indicators can be applied, the mode for deriving their central tendency and the diversity for their dispersion as expressed below

Municipality	Mode	Diversity
A	3	4
B	4	2
C	2	3
D	2	2
E	1 / 2	4
F	3	3
G	4	3
H	3	2
I	1	2

Statistical descriptors applied to summarise the change in number of inhabitant during the period 1900 – 1990

EXERCISE

- Municipalities A, F and H have the same modal value, In what respect their diversity value contributes to express a difference in change behaviour?
- What interpretation can be made about the municipality E based on its modal property (bi-modality) and its diversity?

The time series describing the number of inhabitant during the period 1900-1990 for 9 municipalities are quantitative data measured at cardinal level, therefore six different indicators can be applied for deriving their central tendency and dispersion

Municipality	Median	Interquartile	Mean	Std. Dev.	Coef. Var.	Range
A	945	336.25	1052.7	451.340	0.429	1503
B	4096.5	1046.5	4354.1	967.618	0.222	3039
C	562	39.75	553.4	46.184	0.083	164
D	769	331.25	862	242.390	0.281	672
E	425.5	688.25	706.8	499.151	0.706	1389
F	1200	2858.75	2341.1	2160.560	0.923	5809
G	1516.5	602	1619.3	444.721	0.275	1352
H	1403.5	3482.25	2825.9	2523.657	0.893	7027
I	81	27.75	73.6	18.362	0.249	50

Statistical descriptors applied to summarise the change in number of inhabitant during the period 1900 – 1990

Change trend modelling

The objective of such modelling is to summarise with a single index value the time change behaviour of each observation. Such a constraint limits the panel of functions to very simple ones. Two models expressing the trend of change could be retained: the *linear regression* [5] and the *allometric function* [6]. More precisely, a single coefficient for each model will be retained in order to express the change trend.

The trend of a linear model can be summarised as the slope coefficient *b* of the linear regression function Y = a + *b* X. In the case of time change modelling, the variable X is the time and Y is the considered phenomenon. In a growth process, as in many change processes, the property values of a phenomenon do not often increase in a linear manner with respect to time. Time series values can be transformed by the application of a logarithmic (*log*) or a square root (*sqrt*) function in order to compensate for this non linear increase. As the linear trend coefficient should describe best the change behaviour, the analyst should apply to the original time series values the most efficient transformation to compensate for this non linear behaviour.

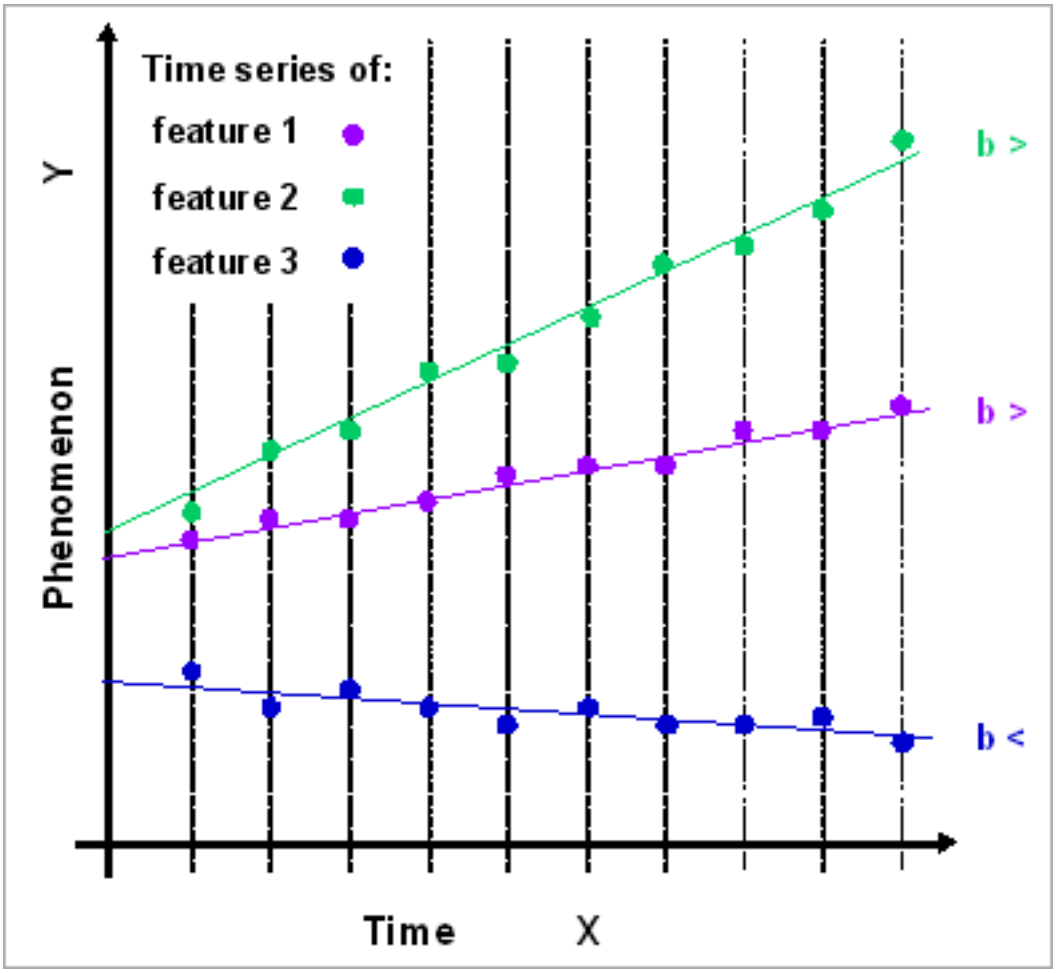
This linear model can then take the following forms:

- For a **linear increase** in the original time series values: Y = a + b X
- For a **non linear increase** in the original time series values: *log*Y = a + b X or *sqrt*Y = a + b X
  with X as the time variable and Y the considered phenomenon

The principle of modelling the evolution trend is illustrated below. From the three time series illustrating property change of three hypothetical situations throughout time, one can observe that the slope coefficient *b* of a linear regression function is capable to summarise the evolution pattern of each series.

[5] A regression function that relates a dependant variable Y with one or several independant variables Xi in a linear manner. A first degree polynomial function is a linear function (see Polynomial regression function)

[6] A regression function that describes the growth rate of a part with respect to the growth of the entire organism (see Allometry)



Modelling the behaviour of 3 time series with the use of their respective slope coefficient b of the linear regression

Assuming the adjustment of the linear trend is satisfying for each individual time series, then b coefficient can be interpreted as the growth rate. The coefficient value expresses the strength of change, a positive value indicates an increase as a negative value suggests a decrease throughout the period of time. However the comparison between coefficient values is limited due to the influence of the size of values on the trend coefficient b. There are several techniques proposed to standardise this coefficient for comparison purposes, but an efficient way to perform comparison is the use of an allometric function.

Allometry [7] is a model developed in the context of biology. It attempts to describe the relative growth of a body part with respect to the growth of the whole body (organism). With the rise of the systemic approach in sciences, this concept of relative growth was then associated with principles of interactions between a set and its subsets (parts). Let us start with a definition: *"Allometry: the relative growth of a part in relation to an entire organism or to a standard"* (Anonymous). Thus one can summarise the relative growth of individual features with respect to the growth of the whole, in our case the set of considered features.

The allometric law is defined as follow:

- In original form:

[7] Allometry is a concept developed in biology. "Allometry: the relative growth of a part in relation to an entire organism or to a standard" (Merriam-Webster)

$$P_i = a + O^b$$

with:

P_i: size of part i

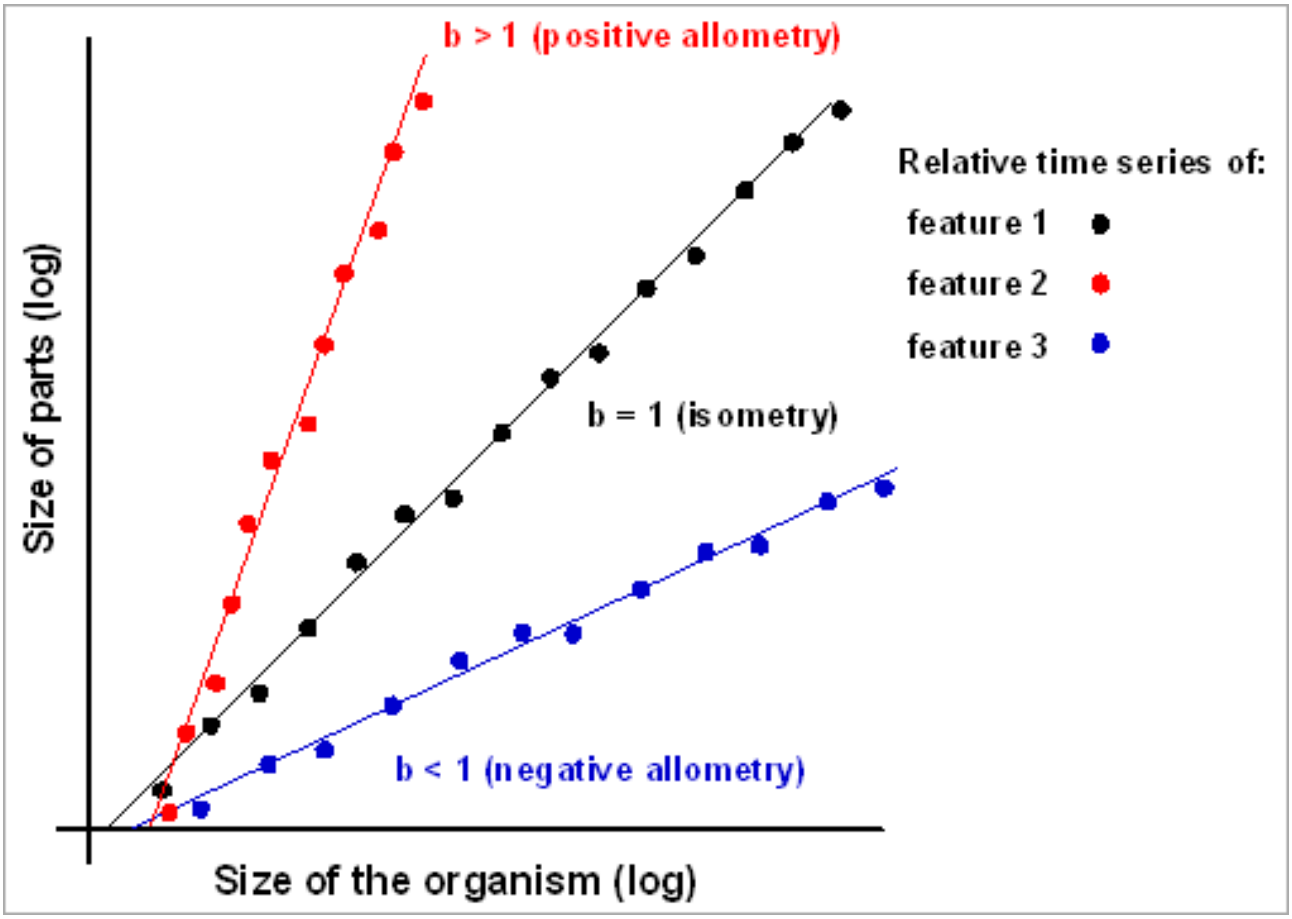
O: size of the entire organism

b : allometric coefficient

- In linear form:

$$log\ P_i = log\ a + b\ log\ O$$

In this linear form the allometric coefficient b -which is also the slope coefficient of the linear regression model- can be interpreted as follow:



Modelling the growth of 3 features throughout time with respect to the growth of a whole. The b allometric coefficient (slope) can be interpreted as the relative growth rate of the corresponding feature.

The b change index provides information about the type and the rate of relative change of individual features. Comparison between bi index values is then made possible, as well as mapping their spatial distribution within the study area.

Illustration

Let us now illustrate the use of *regression slope coefficient* and *allometric coefficient* to summarise the evolution of telephone subscriptions in each municipality between 1900 and 1990. We know that individual evolution is influenced not only by factors governing the diffusion of innovation, but also by the increase of potential subscribers (households) within the considered period of time.

Change in number of telephone subscription during the period 1900 – 1990 with an interval of 10 years

The growth trend of each municipality during this period of time can be first summarised with the use of the *slope coefficient b*. As growth rates are not linear (**Figure**) it is necessary to apply a transformation to original data. In this case the most adequate transformation is the square root function (Sqrt). Transformed data can now be fitted with a linear regression model:

*sqrt*Y = a + b X

with X as the time variable and Y the number of telephone subscription for each municipality

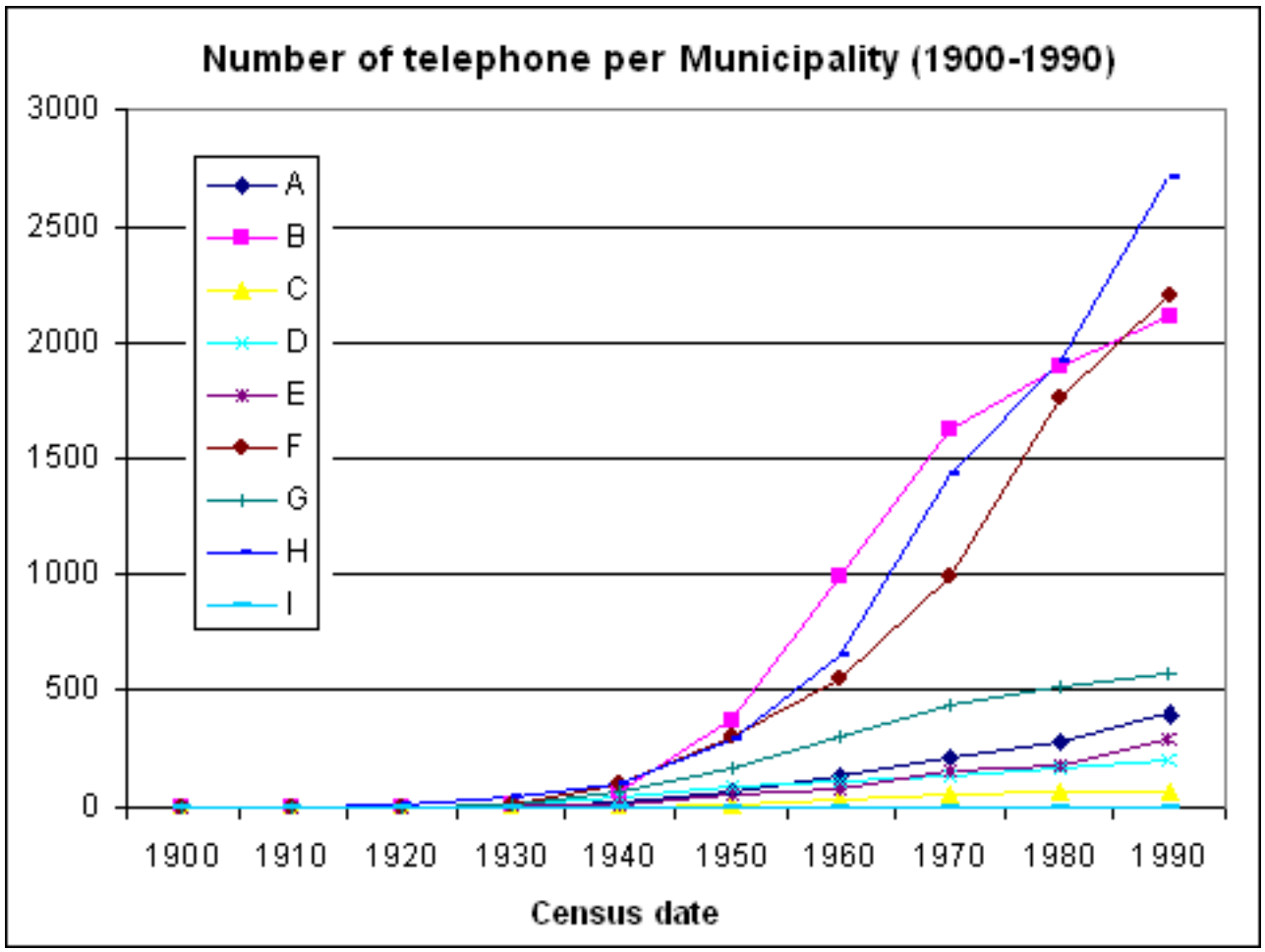


Diagram showing the evolution of number of telephone subscription during the period 1900-1990 for each municipality. It clearly illustrates the non linear progression for all features.

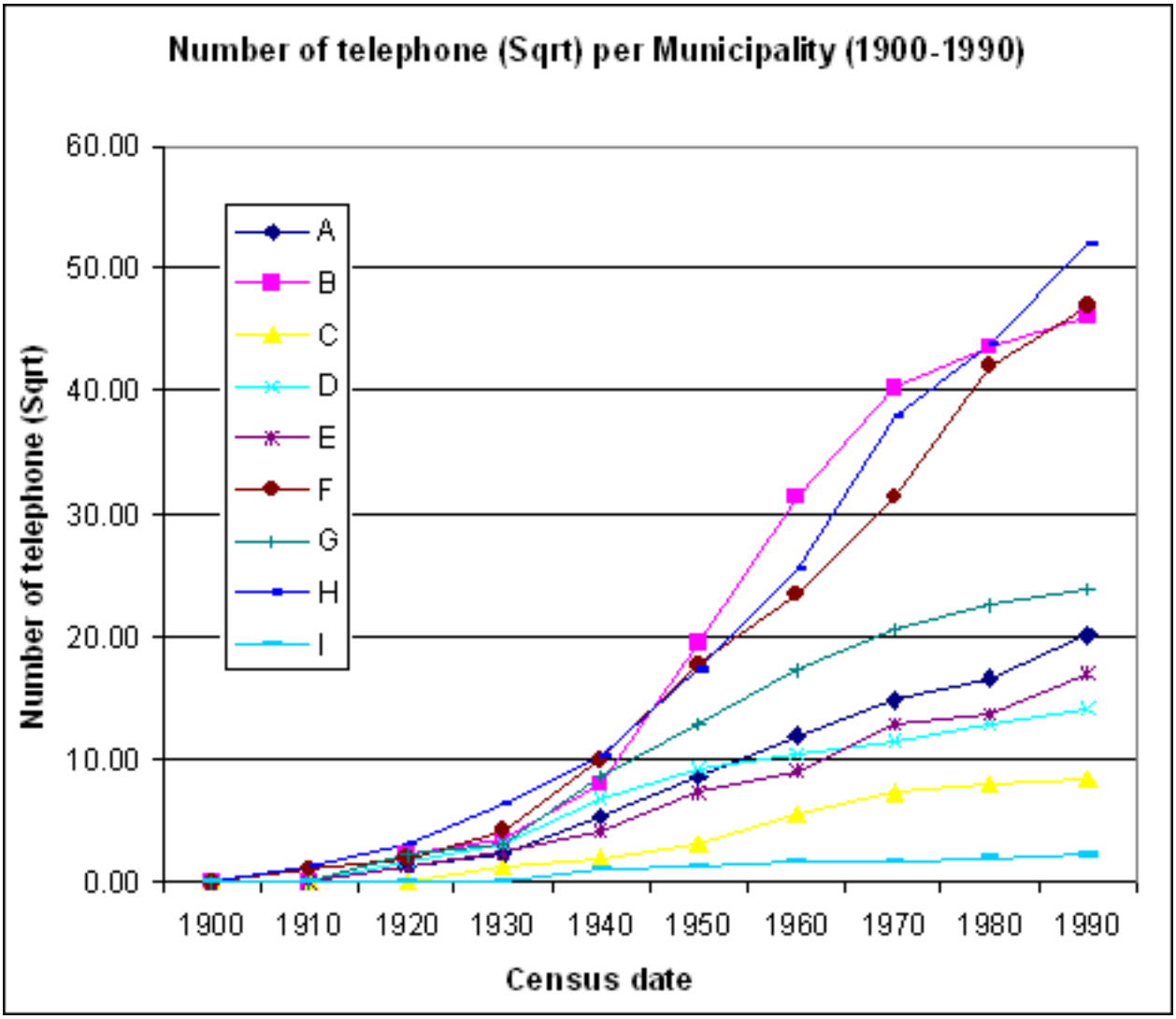


Diagram showing the effect of the square root transformation applied to original data illustrated in the previous figure.

Individual growth rate is summarised in the following table with the use of their respective *b* slope coefficient. Because this process has started only around 1920 for most municipalities the slope coefficient was calculated also for the period 1920-1990.

Municipality	slope 00-90	slope 20-90
A	0.240	0.279
B	0.609	0.733
C	0.110	0.131
D	0.177	0.181
E	0.200	0.231
F	0.558	0.683
G	0.311	0.346
H	0.608	0.740
I	0.029	0.034
District	1.134	1.355

Slope coefficients calculated for the whole period 1900-1990 as well as for the period 1920-1990 on square root transformed data.

Several comments can be made about these coefficients:

- The slope coefficient expresses the absolute amount of subscription increase during the considered period. This is confirmed by the strong correlation between the number of telephone subscription in 1990 and the slope coefficient value.
- Such index values should be interpreted as a global rate of increase during the whole period. Nothing is said about the relative dynamics among features.

EXERCISE

Compare the change trend index (slope coefficient values) for the 2 considered periods 1900-1990 and 1920-1990:

- What general considerations can be made?
- In the following Table, rank municipalities according to their respective slope values for the two periods. Compare and comment on rank changes:

Rank	Municipality Slope 1900-1990	Municipality Slope 1920-1990
1		
2		
3		
4		
5		
6		
7		
8		
9		

Rank and compare the number of telephone subscriptions in 1990 (**Table**) and the slope values for the two periods:

Rank	Municipality Subscriptions 1990	Municipality Slope 1900-1990	Municipality Slope 1920-1990
1			
2			
3			
4			
5			
6			
7			
8			
9			

Let us now illustrate the contribution of the allometric coefficient as a change index. As you remember the linear form of the allometric function is the following:

$$log\ P_i = log\ a + b\ log\ O$$

with:

P_i: size of part i

O: size of the entire organism

b : allometric coefficient

The allometric coefficient is therefore the slope coefficient b of the function relating the growth of a part P_i (here the Municipality i) with the one of the whole O (here the District).

Individual relative growth rate is summarised below with the use of their respective *b* allometric coefficient. Because this process has started only around 1920 for most municipalities the allometric coefficient was calculated also for the period 1920-1990.

Municipality	1900-1990	1920-1990
A	0.716	0.935
B	0.940	1.151
C	0.516	0.792
D	0.644	0.716
E	0.663	0.854
F	0.911	1.088
G	0.775	0.883
H	0.874	0.970
I	0.171	0.295
District	1.000	1.000

Allometric coefficients calculated for the whole period 1900-1990 as well as for the period 1920-1990.

Several comments can be made about these coefficients:

- The allometric coefficient expresses the relative growth rate of each municipality during the considered period. This is confirmed by the coefficient value of 1 for the district which corresponds to the entire set of municipalities.
- The sub-period 1920-1990 expresses best the relative growth rates, as most of municipalities have started only from 1920.
- For the period 1920-1990 there are two cases of positive allometry and seven cases of negative allometry. However municipality H is very close to isometry. Municipality B has the highest coefficient value while the municipality I has from far the lowest.

EXERCISE

Compare the change index (allometric coefficient values) for the 2 considered periods 1900-1990 and 1920-1990:

- What general considerations can be made?

- In the following Table, rank municipalities according to their respective coefficient values for the two periods. Compare and comment on rank changes:

Rank	Municipality Allom. 1900-1990	Municipality Allom. 1920-1990
1		
2		
3		
4		
5		
6		
7		
8		
9		

Rank and compare the number of telephone subscriptions in 1990 (**Table**) and the allometric coefficient values for the two periods:

Rank	Municipality Subscriptions 1990	Municipality Allom. 1900-1990	Municipality Allom. 1920-1990
1			
2			
3			
4			
5			
6			
7			
8			
9			

Then compare the observations made for the slope coefficients and the allometric coefficients.

Global property change:

Another way to summarise the behaviour of individual features throughout time is to synthesise the time dimension variables into synthetic component. Finally index values will be attached to each feature (component scores) but they are derived from a global analysis of the whole set of features and time variables. The objective

is again to summarise the individual time change behaviour with a couple index values. Such a transformation is called *Principal component transformation or analysis (PCA)* [8], it produces *principal component scores* for each feature.

Standardised principal component scores

One can read the content of the following table as a set of time series describing the change in number of inhabitant during a period of time for the nine considered municipalities (row direction reading). This table structure can be approached from the column direction as well. This time census dates are interpreted as *time variables*. Each of them describes the number of inhabitant in the nine municipalities at a specific date. As properties change smoothly from a census date to another, one can expect a fairly high degree of correlation between time variables. Therefore it seems reasonable to synthesise this set of variables (here 10 dates or 10 variables) into a few number of relevant components.

Municipality	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990
A	615	701	757	821	921	969	969	1156	1500	2118
B	3453	3956	4160	3259	3634	4033	4248	4932	5568	6298
C	523	561	546	600	576	563	448	532	573	612
D	683	661	650	690	768	770	812	1084	1180	1322
E	311	328	339	354	363	488	761	1303	1121	1700
F	774	899	660	822	1066	1334	1813	4329	5245	6469
G	964	1280	1267	1436	1463	1570	1621	2021	2255	2316
H	856	907	1021	1234	1317	1490	2584	5214	5753	7883
I	87	95	81	83	95	81	52	58	45	59

Change in number of inhabitant during the period 1900 – 1990 with an interval of 10 years. Features are the nine municipalities and variables are the 10 census dates.

This high degree of correlation is illustrated by the corresponding correlation matrix in the next **table**. One can observe the following:

- The strongest correlation takes place between a date and its preceding or its next one.
- The degree of correlation decreases with the distancing of census dates.
- The variable 1960 has the overall highest degree of correlation with all other variables.

[8] A procedure that transforms an original set of variables into a set of Principal components. This transformation removes the original correlation between variables (information redundancy) and structure the overall variability into ordered components (the first component carrying more variability than the second, and so on)

1900	*1*	.998	.996	.979	.985	.981	.906	.636	.631	.542
1910	.998	*1*	.997	.984	.988	.985	.904	.630	.626	.534
1920	.996	.997	*1*	.985	.986	.979	.901	.613	.605	.518
1930	.979	.984	.985	*1*	.997	.991	.937	.686	.680	.596
1940	.985	.988	.986	.997	*1*	.998	.946	.709	.707	.623
1950	.981	.985	.979	.991	.998	*1*	.960	.743	.740	.659
1960	.906	.904	.901	.937	.946	.960	*1*	.891	.881	.831
1970	.636	.630	.613	.686	.709	.743	.891	*1*	.996	.989
1980	.631	.626	.605	.680	.707	.740	.881	.996	*1*	.990
1990	.542	.534	.518	.596	.623	.659	.831	.989	.990	*1*

Correlation matrix for the ten time variables. It shows the strong degree of correlation between variables.

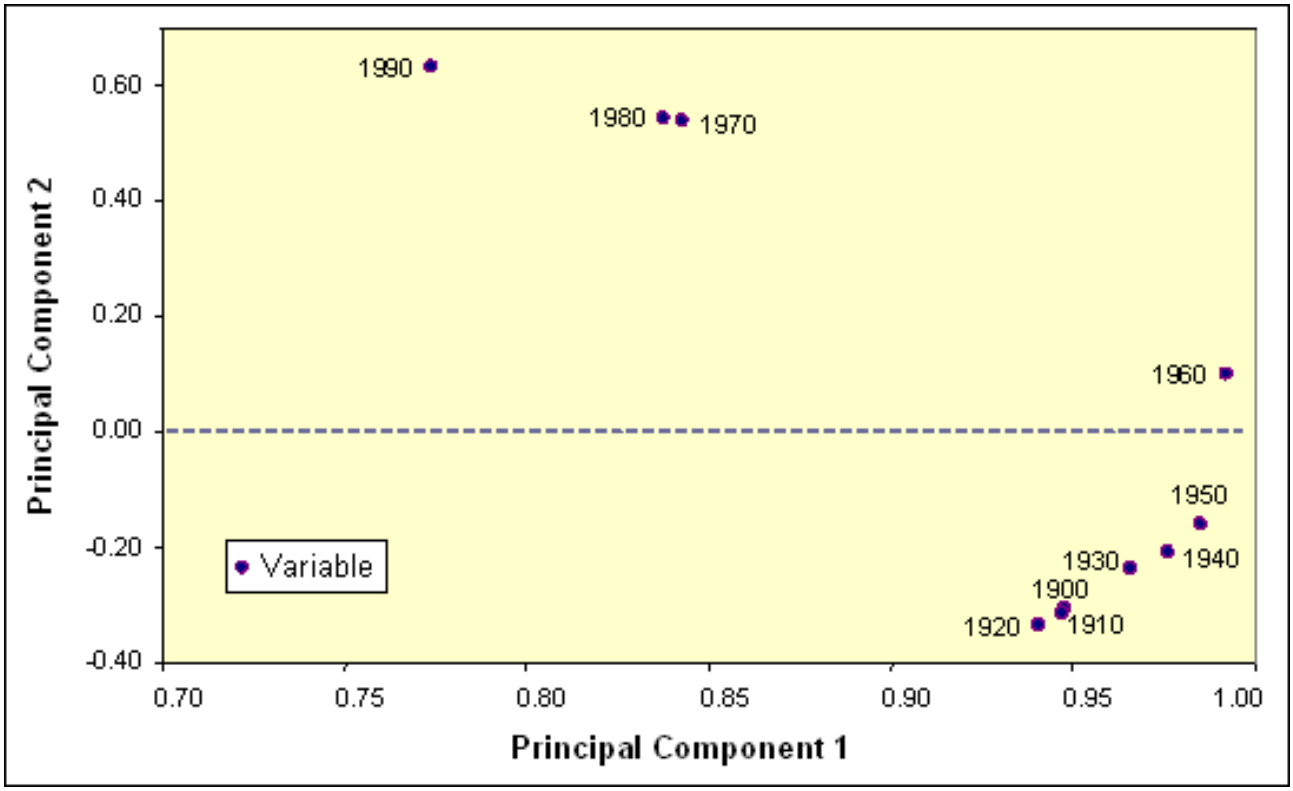
In order to remove the undesirable scale effect, variables should be standardised prior to the component transformation.

The principal component transformation produces two significant components, summing up 99.5% of the total variance:

- component 1 (PC1): 85.3% of variance
- component 2 (PC2): 14.2% of variance

The contribution of original variables to the two retained principal components is expressed by their respective weights in the component matrix. The figure below illustrates this contribution. One can observe the following:

- Years 1900 to 1960 have a strong influence on the component 1 (1960 with the strongest), while 1970 to 1990 have a lower one (1990 with the weakest). All weights are positive.
- For the component 2 weights range from negative to positive. Once again the 2 groups of variables can be identified. The order of weights almost follows the sequence of years.



Weight of variables on the two principal components.

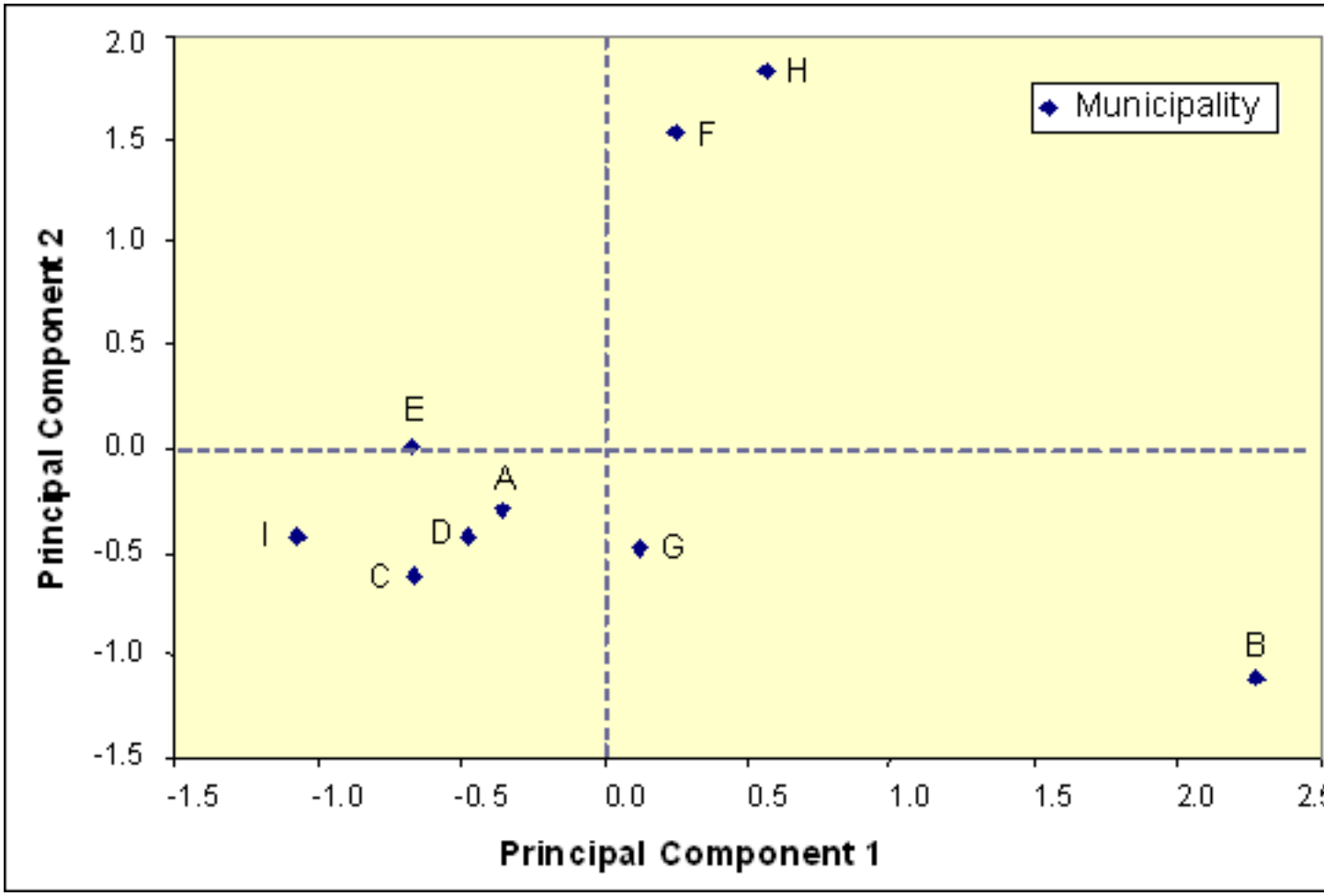
EXERCISE

Assign a name to the two principal components that describes best their content:

- Component 1: (relative population size during the period)
- Component 2: (recent growth rate versus early one)

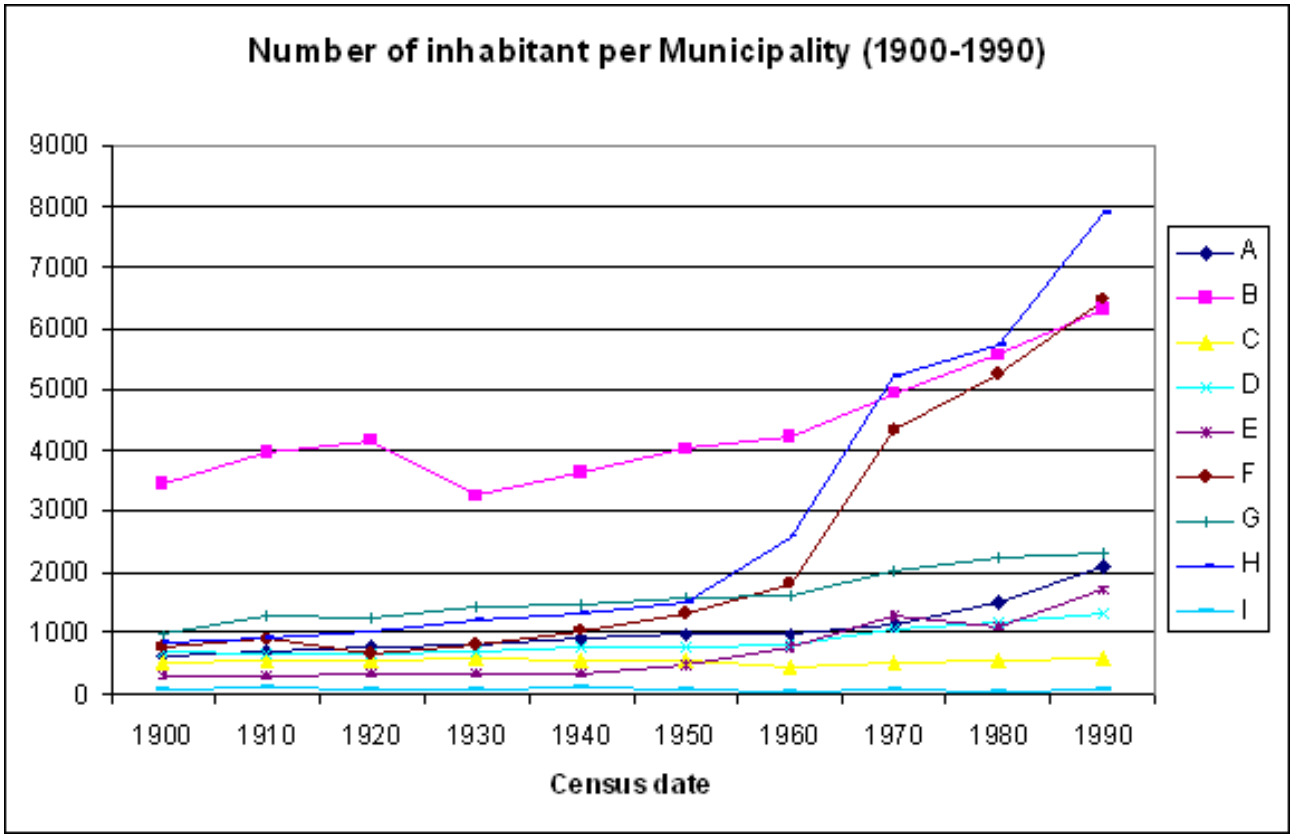
Now let's consider the properties of the 9 municipalities for the two principal components. Their respective score values can be plotted for identification of groups with similar evolution behaviour. From the next **figure** three groups of features can be identified:

- **Group 1**: made of a single outsider B
- **Group 2**: made of municipalities F and H. Their two index values are positive
- **Group 3**: the largest group that includes the 6 remaining municipalities. Index values (scores) for both components are almost negative. However, municipality G is slightly away from other members of this group, with a positive index value for the first component.



Score values of the 9 municipalities for the two principal components.

Interpretation of score values as a global behaviour index is made possible by comparing with the diagram of population change during this period of time.



Population change for the 9 municipalities during the period 1900-1990.

EXERCISE

Compare curve shapes of municipalities from the last diagram with their respective score values plotted in the last figure. Characterise and comment on the proposed grouping:

- Group 1:
- Group 2:
- Group 3:

1.2. Time series behaviour description

How to describe the sequence of changes of each individual feature?
Let us now concentrate on the individual change pattern of features. Information about each feature consists of a sequence of properties called *Time series*. Information content of a time series should be carefully considered for the selection of appropriate analysis tools. In the thematic dimension first, as seen before, properties can be measured at nominal level (qualitative data), at ordinal or cardinal level (quantitative data).
Furthermore information content is also concerned by the time dimension:

- At ordinal level: only the order of the sequence is considered. In other words intervals of time (or spacing) are discarded.
- At cardinal level: intervals of time are significant for the analysis.

Originally data can be collected at regular but also at irregular intervals of time. As most techniques for time series analysis require regular intervals, equal spacing procedures should be applied to the original time series prior its analysis.
The following table presents a selection of methods for the analysis of individual time series.

	TIME DIMENSION **Intervals**
UNIVARIATE **Single Observation (Time series)**	**Detailed description:** Runs test (Qual) Markov chains / Transition matrices (Qual) Auto-association (Qual) Auto-correlation (Quant) Filtering (Qual/Quant) Fourier series (Quant) Time regression function (Quant) Allometry (Quant)

Qual: qualitative data (nominal) Quant: quantitative data (ordinal, cardinal)

A brief overview of time series analysis methods

1.2.1. Time as a sequence of events (with regular intervals)

In this context the behaviour of each feature is expressed throughout the considered period of time as a succession of properties or events measured during an interval of time ti. In order to illustrate several of the methods discussed in this Section, let us take a simple example. We are interested in the analysis of car accidents occurring in the municipalities of our study area during a period of 3 years (1988 to 1990). The considered phenomenon is then the frequency of car accidents reported to the police during each month in this period in each of the municipalities. From these reports a time series made of 36 monthly counts was then produced. It describes the behaviour or the profile of each municipality during this 3 years period of time. Let us now concentrate on the two municipalities E and I with their respective time series listed in the table below.

Month	Municipality E	Municipality I
Jan.88	1	4
Feb.88	0	3
Mar.88	2	2
Apr.88	1	2
May.88	2	0
Jun.88	4	2
Jul.88	6	4
Aug.88	3	3
Sep.88	3	2
Oct.88	2	1
Nov.88	1	1
Dec.88	2	3
Jan.89	0	6
Feb.89	0	2
Mar.89	2	1
Apr.89	1	0
May.89	2	1
Jun.89	3	2
Jul.89	5	4
Aug.89	4	3
Sep.89	3	0
Oct.89	2	1
Nov.89	1	0
Dec.89	1	2
Jan.90	2	3
Feb.90	1	2
Mar.90	0	1
Apr.90	1	0
May.90	2	1
Jun.90	3	2
Jul.90	4	5
Aug.90	3	3
Sep.90	3	1
Oct.90	2	1
Nov.90	1	2
Dec.90	0	3

Number of car accidents per month recorded during the period 1988 to 1990 for the two municipalities E and I

The behaviour of individual features can be described and analysed in many different ways with various considerations about the time dimension:

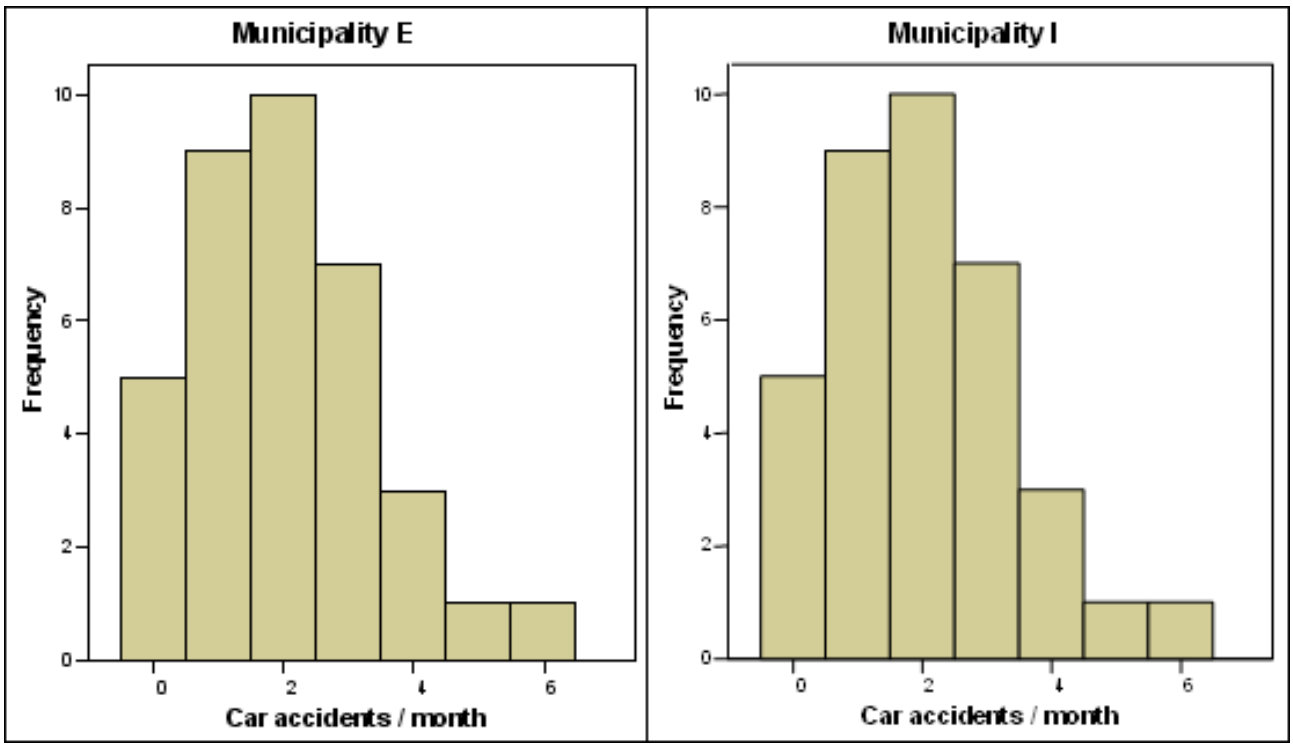
- In which way properties change during the period of time?
- Is there any regular scheme of properties?
- How this individual behaviour compare with a reference behaviour?

- Can this behaviour be generalised and modelled?

Let us first briefly summarised the overall property distribution for the two municipalities during this 3 years period of time. Descriptive statistics indices provide us with basic characteristics about property values present in each time series. Numerical indices listed in the table below indicate that the total number of car accidents is 73 for both municipalities and that their respective mode, mean, median, and standard deviation are identical. Their histogram shown in the next figure is confirming an identical distribution of property values during this period of time.

Statistics	Municipality E	Municipality I
N	36	36
Mode	2	2
Median	2	2
Mean	2.03	2.03
Minimum	0	0
Maximum	6	6
Std. Deviation	1.444	1.444
Skewness	0.67	0.67
Kurtosis	0.42	0.42
Total Nb. of car accidents	73	73

Descriptive statistics for the two time series E and I



Histogram distribution for the two time series E and I

Furthermore the two distributions follow almost perfectly a Poisson distribution with a mean value of 2.03. A Chi-square test as well as a Kolmogorov-Smirnov test indicate that the differences in frequency distribution between the Poisson distribution and the two observed distributions are strongly not significant. This indicates

that during the period of 36 months the frequency of monthly car accidents do not depart from a random distribution having a mean value of 2.03. In other words, there is no individual factor in both municipalities that strongly affects the variations in the frequency of monthly car accidents. The following table shows the comparison between a Poisson frequency distribution and the two observed distributions.

Observed distributions		Poisson distribution		Difference
Nb. of car accidents per month (k)	Frequency for 36 months	Probability with $\mu = 2.03$	Frequency for 36 months	Frequency difference
0	5	0.135	4.90	0.1
1	9	0.270	9.70	-0.7
2	10	0.271	9.80	0.2
3	7	0.180	6.50	0.5
4	3	0.090	3.00	0.0
5	1	0.036	1.00	0.0
6	1	0.012	0.40	0.6

Frequency distribution of monthly car accidents in municipalities E and I compared with a Poisson distribution having the same mean value.

Can we then conclude to an identical behaviour of this phenomenon in the two municipalities during this 3 years period? Let us observe the contribution of the time dimension for an in depth analysis of individual feature behaviour.

Runs test:

Let us analyse the sequence of properties occurring within a time series. Our interest is in the *succession pattern* of property values. At ordinal or cardinal level property values can increase or decrease regularly or can present a variety of change patterns. For a measurement at nominal level our interest is in the change of a category to another one.

A *runs test* [9] operates at nominal level and more precisely for a binary variable with only two possible properties or states. One should therefore accommodate the original time series by a transformation grouping all possible properties into two states 0 and 1.

A runs test aims to compare the observed time series with a random sequence of states. It is used to test for randomness of occurrence. Let us consider the experiment of tossing a coin and the time series as the result of 16 successive tosses. Assuming an equal probability of 0.5 for obtaining a head (H) or a tail (T) for each toss, a large variety of sequences combining 8 heads and 8 tails can be obtained. The two following sequences illustrates extreme situations that unlikely occur at random:

- Grouped sequence: H H H H H H H H T T T T T T T T
- Regular alternation: H T H T H T H T H T H T H T H T

[9] A runs test aims to compare an observed time series with a random sequence of states

The succession of states corresponds to a specific pattern of the considered time series. In order to describe the pattern the full sequence is subdivided into runs. *A Run* is defined as an uninterrupted succession of the same state. In our previous example only 2 runs can be identified for the grouped sequence as 16 runs occur in the regular alternation sequence. For a random sequence the number of runs should be situated between these two extreme values.

It is admitted, when the number of occurrence n_1 and n_2 for each of the two states exceeds ten, that the distribution of random arrangement of two states within a sequence can be approximated by a normal distribution with an expected mean Û and its variance #^{2}û defined as follow:

$$\bar{U} = (2n_1n_2 / (n_1 + n_2)) - 1$$

$$\sigma^2_{\bar{U}} = (2n_1n_2 (2n_1n_2 - n_1 - n_2)) / ((n_1 + n_2)^2(n_1 + n_2 - 1))$$

with:

$\bar{U}$: mean number of runs in a random distribution

$\sigma^2_{\bar{U}}$: variance of the mean number of runs in a random distribution

n_1 : number of occurrence for the state 1

n_2 : number of occurrence for the state 2

We can then applied a Z test to compare the observed number of runs U with the expected one from a random sequence:

$$Z = (U - \bar{U}) / \sigma_{\bar{U}}$$

The null hypothesis and its alternative are:

$$H_0 : U = \bar{U}$$
$$H_1 : U \neq \bar{U}$$

With a 5% level of confidence the Z value should be less than –1.96 or greater than 1.96 to reject H0 and then to conclude that number of runs in the sequence is significantly different from the one in a random sequence.

Let us now look at the application of the Runs test to our illustrative examples: first the succession of mayor gender for municipalities E and F and then the sequence of monthly car accidents in municipalities E and I.

The first variable “gender of the municipality mayor” illustrate a series of binary properties, female or male. The next table lists the sequence for each municipality between 1900 and 1990. Thus the original values can be used to identify the number of runs in each sequence for the municipalities E and F. The two sequences are the following, with 1 for Female and 2 for Male:

Municipality E: *2 2 2 2 2 2 2* 1 1 1

Municipality F: *2 2* 1 1 *2* 1 *2 2* 1 *2*

For the municipality E we can count only 2 runs and calculate a value Z = -2.209. Thus the null hypothesis of a random sequence can be rejected with a level of confidence of 5% as the Z value is less than –1.96. The presence of only two runs within the sequence has only very little chance to result from a random arrangement;

therefore there are certainly specific factors that contribute to this situation. On the contrary the sequence for municipality F cannot be considered as significantly different from a random sequence. It contains 7 runs and its Z value 0.492 lies within the critical region (± 1.96).

Year	A	B	C	D	E	F	G	H	I
1900	2	2	2	2	2	2	2	2	2
1910	2	2	2	2	2	2	2	1	2
1920	2	1	2	2	2	1	1	1	2
1930	2	2	2	2	2	1	2	2	2
1940	2	2	2	2	2	2	2	1	2
1950	2	1	2	2	2	1	2	2	2
1960	1	1	2	2	2	2	1	2	2
1970	2	1	2	2	1	2	2	1	2
1980	1	2	2	2	1	1	1	2	2
1990	2	1	1	2	1	2	2	1	1

1: woman as mayor 2: man as mayor

Gender of the mayor for the 9 municipalities during the period 1900-1990.

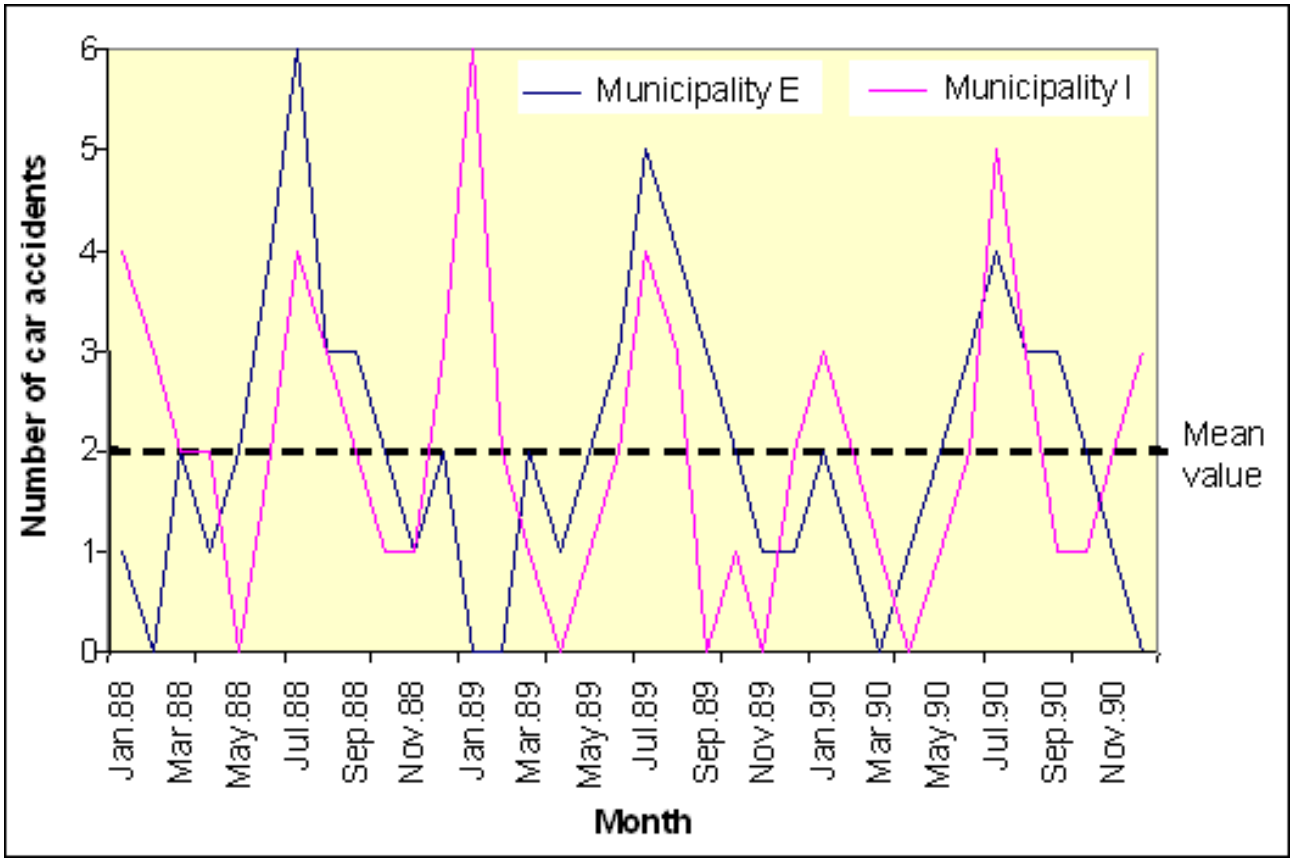
We are now concerned with the description of time series related with the number of monthly car accidents in municipalities E and I (**Table**). We would like to analyse the succession of monthly accidents with respect to the central tendency of the considered time series. Let us then group monthly frequencies into two categories: category 0 containing months with frequencies below the mean value (i.e. 2.03 for both time series) and category 1 including months with frequencies above it.

The following table summarises results for the grouping procedure and for the runs test applied to the time series of municipalities E and I. While both time series have the same number of cases below and above the mean value, their sequential distribution within the same period of time is obviously and significantly different. The number of runs for time series E is significantly less than a random time series distribution, but this is not true for the time series I with a Z value that belongs to the interval of confidence of 5%.

Statistics	Municipality E	Municipality I
Test Value	2.03	2.03
Cases < Test Value	24	24
Cases >= Test Value	12	12
Total Cases	36	36
Number of Runs	7	13
Z	-3.628	-1.337
Asymp. Sig. (2-tailed)	0.000	0.181

Runs test applied to time series of monthly car accidents for municipalities E and I. The threshold or test value is set to the mean value for the grouping of original monthly frequency into two categories.

The graphical representation of the two series in next figure illustrates the differences of time distribution that is pointed out by the runs test. It shows that time series I crosses the threshold value almost twice as much as time series E. One can notice the strong influence of the threshold value on the number of produced runs.



Distribution of monthly car accidents in the municipalities E and I during the period 1900-1990. The threshold assigned for the runs test is the mean value 2.03

When transforming the original time series from an ordinal or cardinal level down to the nominal binary level, different criteria can be applied to determine the threshold value for grouping into two categories:

- Differentiating between values *below* or *above* some threshold (ie. the central tendency as illustrated)
- Differentiating between *"increasing"* and *"decreasing"* situations.

It should be noted that the runs test reports only on the number of runs within the sequence, there is no specific information about the length of each run.

EXERCISE

From the table describing the gender of municipality mayors during the period 1900-1990 (**Table**), calculate the number of runs, the Z value and apply the test for evaluating the type of sequence for municipalities B and D:

- Comment on test conclusions
- Identify the specific situation of the time series for the municipality D.

From the table describing the political majority of municipalities during the period 1900-1990 (**Table**), transform original values into binary properties by choosing a relevant criterion for grouping the four categories into two categories. Then calculate the number of runs, the Z value and apply the test for evaluating the type of sequence for municipalities A and I:

- Comment on test conclusions
- What is the influence of the grouping criterion upon the test and the objective of the test?

Table with transformed value properties

Municipality	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990
A										
B										
C										
D										
E										
F										
G										
H										
I										

Markov chains:

You will recall from the section **Global Property Change** that we were interested in the summary of change within a period of time with the use of only two time markers or limits. Transition matrices were used to describe the global change of political majority in the nine municipalities between 1900 and 1990. We have then introduced the notions of *transition frequency matrix, transition relative frequency matrix* and *transition proportion matrix*. This was applied to summarise the overall change trend within the set of municipalities.

We would like now to analyse the succession of states within a single time series in order to evaluate the probability of transition from one state to another. This refers to a *transition probability matrix* [10]. Furthermore this matrix expresses the probability that a state A will follow a state B, provided B occurs. This is called *conditional probabilities* that are contained in the transition probability matrix.

In complement to the evaluation of global probability of occurrence from any state to any state based on the analysis of the transition matrix, the construction of *Markov chains* [11] offers further investigations on the sequence of state changes:

- to estimate the probability of occurrence from any original state to any final state after a specific sequence of n steps,
- to estimate the probability of occurrence of each intermediate state in a specific sequence of n steps,
- to compare transition probabilities for the observed sequence with some reference models: deterministic, random, uniform, …

[10] A probability matrix that expresses a transition from one state to another

[11] A technique to estimate the probability of occurrence from any original state to any final state after a specific sequence of n time steps. It makes use of transition matrices

Let suppose a time series composed of 64 observations regularly distributed within a period of time as illustrated in the following table.

1 to 10	11 to 21	22 to 32	33 to 43	44 to 54	55 to 64
Start	C	C	C	C	C
C	C	B	C	A	A
C	C	B	C	A	A
C	C	B	C	A	A
A	A	B	C	C	A
A	A	B	A	C	A
A	C	C	A	B	A
A	C	C	A	C	C
A	B	C	A	C	A
A	A	B	A	A	B
C	C	B	C	B	End

Hypothetical time series made of 64 observations regularly distributed in time illustrating the change between three possible states A, B and C.

There are three possible states labelled A, B and C. As seen before, a 3x3 *transition frequency matrix* can be constructed showing the number of times a given state is succeeded by another.

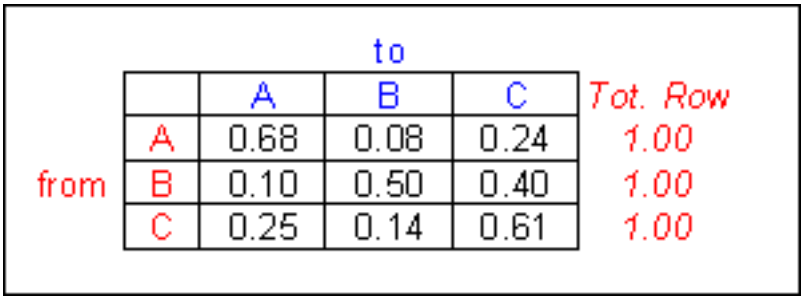
		to A	to B	to C	Tot. Row
from	A	17	2	6	25
	B	1	5	4	10
	C	7	4	17	28
Tot. Col.		25	11	27	63

Transition frequency matrix produced from the sequence of 64 observations. It shows property change patterns

The measured time series contains 64 observations, so there are (n-1) = 63 transitions. Note that the rows and columns totals will be the same, provided the sequence begins and ends with the same state otherwise two rows and two columns will differ by one.
It is then possible to derive the *transition relative frequency matrix* and the *transition proportion matrix* (as expressed in the two following tables)

		to A	to B	to C	Tot. Row
from	A	0.27	0.03	0.10	0.40
	B	0.02	0.08	0.06	0.16
	C	0.11	0.06	0.27	0.44
Tot. Col.		0.40	0.17	0.43	1.00

Transition relative frequency matrix derived from the Transition frequency matrix



from \ to	A	B	C	Tot. Row
A	0.68	0.08	0.24	1.00
B	0.10	0.50	0.40	1.00
C	0.25	0.14	0.61	1.00

Transition proportion matrix derived from the Transition frequency matrix.
It indicates the proportion of succession from any state to any possible state

Comparison with a reference sequence

An observed sequence can then be compared with a reference sequence based on their transition frequency matrix (counts or relative frequency). The reference series can be either a theoretical model (deterministic, random, uniform, …) or another observed sequence. One can use a Chi-squared test to determine if the observed series is significantly different from the reference.

Analysis of the succession of changes

A sequence in which the state at one point is partially dependent on the preceding state is called Markov chains (named after the Russian statistician, A.A. Markov). A sequence having the Markov property is intermediate between deterministic sequences and completely random sequences. “In theory, the probable state of a Markov system at any future time can be predicted from knowledge of the present state” (Davis 1986).

From the transition proportion matrix (**Table**) it is then possible to evaluate the probability of occurrence of any state from any original state after a specific number of time step in the sequence, assuming the relative frequency distribution from the observed sequence is representative from the overall behaviour of the phenomenon. In other words, one assumes that the relative frequencies correspond to the probabilities from the parent population. When handling simple situations with few states and very few steps, it is possible to obtain such probability of occurrence by an experimental approach, but a more general solution can be found with the combination of conditional probabilities.

Let us take a simple situation to illustrate these two approaches. We would like to estimate the probability of ending with the state A when starting with this state A in a two transitions sequence (a sequence with two steps), based on the observed sequence (**Table**).

Experimentally we can start from the transition proportion matrix (**Table**) to construct a diagram of all possible two steps sequences starting from the state A with their corresponding conditional probabilities.

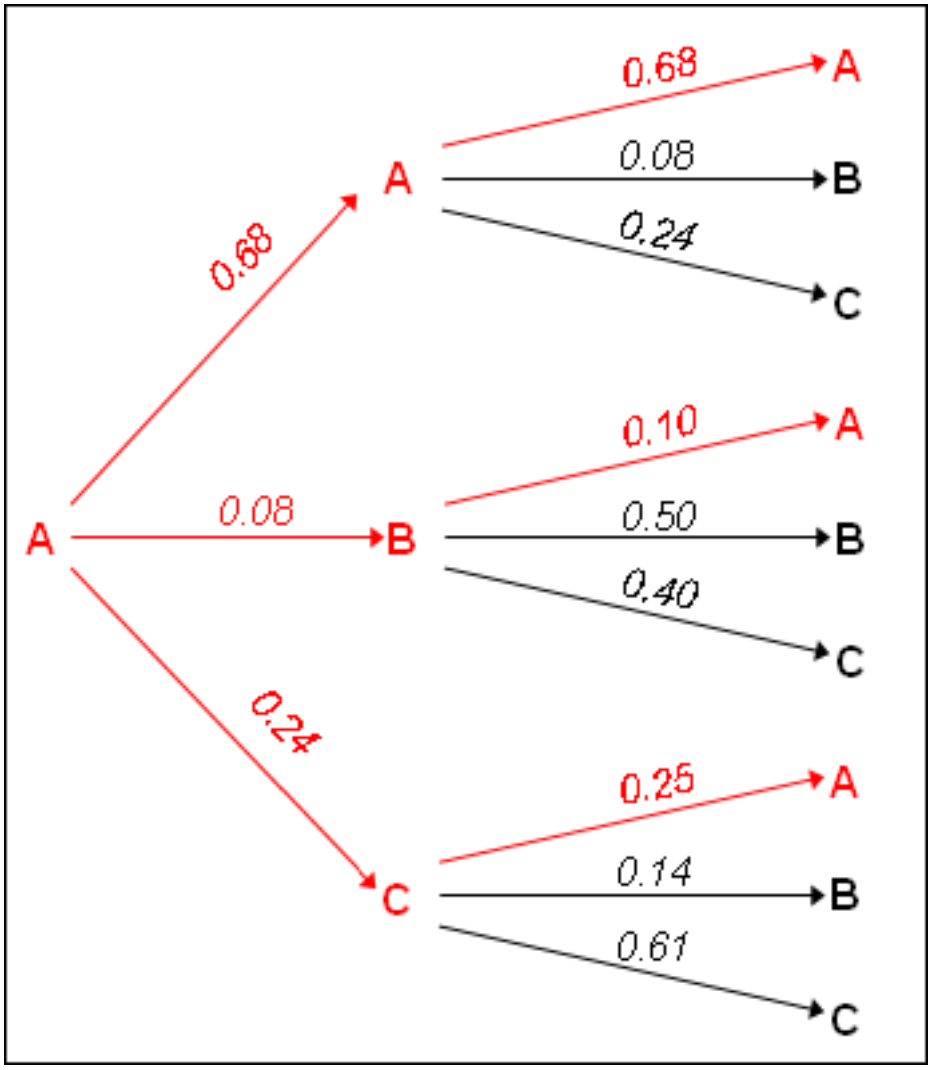


Diagram showing all possible sequences of two steps starting from state A with their corresponding conditional probabilities

From this diagram one can identify three corresponding sequences:

$$A \rightarrow A \rightarrow A,\ A \rightarrow B \rightarrow A \text{ and } A \rightarrow C \rightarrow A$$

The probability of occurrence for each sequence is obtained as following:

- AAA: $Pr_{A \rightarrow A} * Pr_{A \rightarrow A} = 0.68 * 0.68 = 0.462$
- ABA: $Pr_{A \rightarrow B} * Pr_{B \rightarrow A} = 0.08 * 0.10 = 0.008$
- ACA: $Pr_{A \rightarrow C} * Pr_{C \rightarrow A} = 0.24 * 0.25 = 0.060$

Then the overall probability of ending with the state A when starting with this state A in a two transitions sequence is:

- AiA: $Pr_{A \rightarrow i \rightarrow A} = 0.462 + 0.008 + 0.060 = 0.530$

It then become tedious to compute experimentally all other possible sequences of combination of three different states. Furthermore when the number of states and the number of steps increase, this becomes simply impossible to achieve. This can be obtained with ease by matrix algebra.

In order to derive the probability of obtaining any *i* state from any original state after *n* steps, the resulting probability matrix is simply the original "transition proportion matrix" called *probability matrix* [P] powered to the number of step *n*: $[P]^n$.

When applied to our above example, the resulting probability matrix $[P]^2$ illustrated in the following table shows not only the experimentally resulting probability $Pr_{A\#i\#A}$ but also probabilities for all other combinations.

0.68	0.08	0.24	X	0.68	0.08	0.24	=			to		
0.10	0.50	0.40		0.10	0.50	0.40				A	B	C	Tot. Row
0.25	0.14	0.61		0.25	0.14	0.61		from	A	**0.53**	**0.13**	**0.34**	1.00
									B	0.22	0.31	0.47	1.00
									C	0.34	0.18	0.49	1.00
[P]				**[P]**						**$[P]^2$**			

Probabilities of obtaining each 3 states A, B and C from each of the same 3 original states after 2 steps. PrA#i#A is identified in the resulting matrix.

It is interesting to observe that when the number of step becomes important the rows tend to become similar. This indicates that the influence of the original states diminishes with time; it is the expression of the "persistence of memory" in a Markov process, as illustrated below.

0.68	0.08	0.24	0.46	0.15	0.39	0.40	0.18	0.42
0.10	0.50	0.40	0.30	0.24	0.46	0.38	0.19	0.43
0.25	0.14	0.61	0.37	0.18	0.45	0.39	0.18	0.43
1 step: [P]			**3 steps: $[P]^3$**			**9 steps: $[P]^9$**		

Loss of influence of the original states when the number of step increases.

EXERCISE

Complete the following Table with the resulting probabilities attached to states B and C based on the probability diagram (**Figure**).

Pattern	Probability
A → A → A	0.462
A → A → B	
A → A → C	
A → B → A	0.008
A → B → B	
A → B → C	
A → C → A	0.06
A → C → B	
A → C → C	
Total	

Pattern	Computation	Probability
A → i → A	0.462+0.008+0.060	0.530
A → i → B		
A → i → C		

Time dependency:

From our everyday experience we are aware that the current property of a phenomenon is related with its property a moment before as well as a moment after. This influence is known as the *time dependency*. If we consider the physical phenomenon air temperature, we can feel that temperature properties are changing throughout days, months and seasons but in a more or less continuously manner. We can then express the *rate of change* of temperature within a specific period of time. Many physical, but also social and economical phenomena present such a continuous temporal change of properties, although they can include some more or less abrupt discontinuities from time to time. One analytic interest is to describe this rate of change that expresses somehow the strength and duration (length) of the time dependency. At the opposite one can found phenomena with no temporal continuity, they are called *chaotic* as properties are distributed like randomly throughout time. This influence observed in the time dimension can be extended to the geometrical or spatial dimension. The most obvious example is certainly the distribution of elevation along a profile. Elevation properties vary continuously from a location to the contiguous one. This is known as the spatial dependency of the phenomenon. The rate of change and the duration of the dependency can be also estimated in this geometrical dimension.

Another interesting property of change is certainly the identification of *periodicity* [12] or sequences within a period of time. When the periodicity is obvious we are then observing a cyclic phenomenon. Perfect cyclic phenomena are very rare in the reality and this property depends on the time scale considered. The air temperature illustrates once again a cyclic phenomenon with regular periodicities: daily, seasonally, annually, … Although such repetitive pattern is not perfect, it is then possible to identify *similar patterns* within a considered period of time.

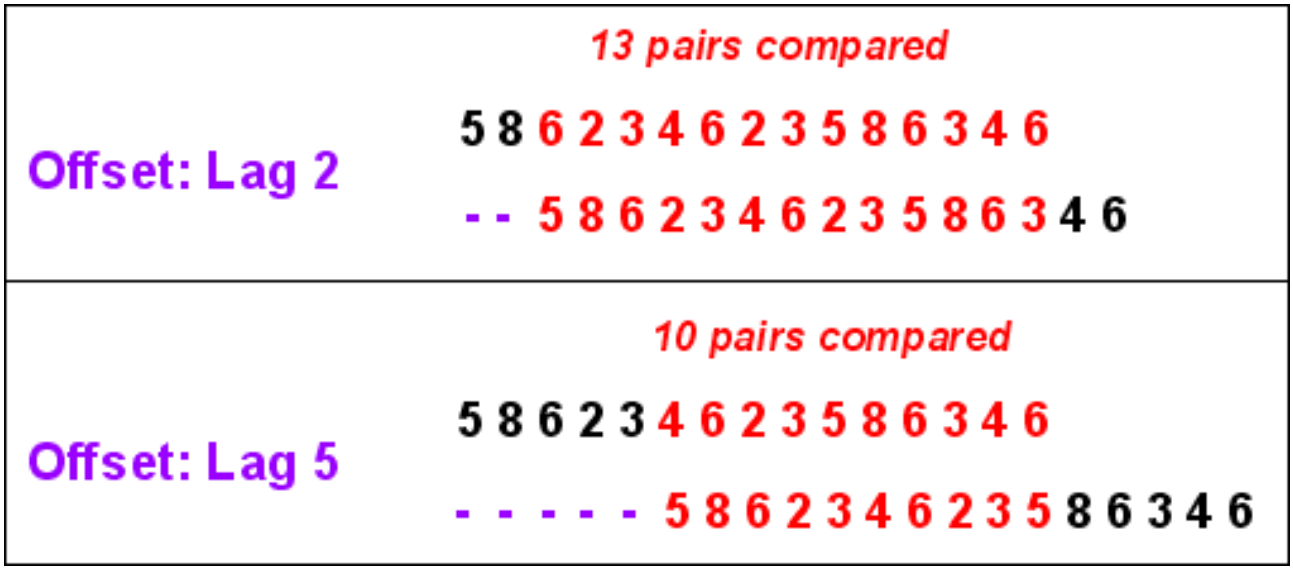
As in the context of time series the description of property change is expressed by regular measures throughout the time period, the indication of similarity can be estimated by a coefficient of correlation. One can then compare the property of each step in the period of time with the one from the next and the following steps successively. This is known as the *auto-correlation technique* [13]. Practically the correlation coefficient is

[12] A succession of properties that occurs regularly throughout time. For example daily temperatures or seasonal unemployement

[13] A procedure to measure the time dependancy of a phenomenon from a time-series compared to itself at different time lags. (see Lag, Auto-correlation coefficient)

computed between a time series and itself with successive offsets between the time positions (intervals). The amount of offset between the two time series is called a *lag* [14]. When the two series are correlated with no offset, the lag equals 0 and of course the correlation is perfect and without any interest. Assuming a time series made of n positions (measures, steps), one can potentially compute correlation with different offsets varying from 0 to n-1 lags. However for the significance of the correlation coefficient the number of compared pairs must be sufficient. This number depends on the size n of the time series and on the lag value. Usually the recommended maximum number of lags is about n/4

The series of correlation coefficients computed for each successive lag can be represented graphically as a *correlogramme* [15]. It can be interpreted to evaluate the duration and the strength of the dependency as well as the presence and the duration of periodicities.



Principle of auto-correlation technique illustrated for two different offsets: 2 lags and 5 lags on an imaginary short time series (overlapped segments are in red colour).

As the measure of correlation should be adapted to the level of measurement of the time series, one can imagine to use the three typical correlation indicators adapted to each level:

- At **cardinal level**: correlation coefficient of Pearson (or Spearman for detecting non linear correlations)
- At **ordinal level**: correlation coefficient of Spearman
- At **nominal level**: association coefficient of Cramer (Cramer's V).

To test the significance of the similarity between the two series at each specific lag, the computed correlation value is compared with the one obtained from random sequence of values.

[14] In time series analysis, variables can be compared synchronously or with a defined time lag. As time series are generally made of regularly distributed intervals of time, this asynchronous comparison corresponds to one or several lag steps

[15] A graphical representation of a succession of correlation values varying throughout time or space (see Lag, Auto/Cross-correlation)

However, in practice we tend to limit the content of time series to two situations: nominal level for categories and cardinal level for continuous values. Specific auto-correlation indicators are developed: the linear *auto-correlation coefficient* [16] for the cardinal level and the *match ratio* for a measure of auto-association at the nominal level.

Auto-correlation:

The linear correlation coefficient calculated at each lag L is the following:

$$r_L = COV(x_i, x_{i+L}) / s^2_X$$

Regardless of the number of pairs considered at each lag value, the denominator of the ratio rL corresponds to the variance of the whole time series. When L=0 then the correlation coefficient corresponds to the linear correlation coefficient of Pearson. Thus r_L value varies between –1 and +1 and can be interpreted as the Pearson's coefficient:

A + sign indicates a direct correlation as a – sign indicates an inverse correlation

The strength of the correlation varies between 0 for no correlation to 1 for a perfect one.

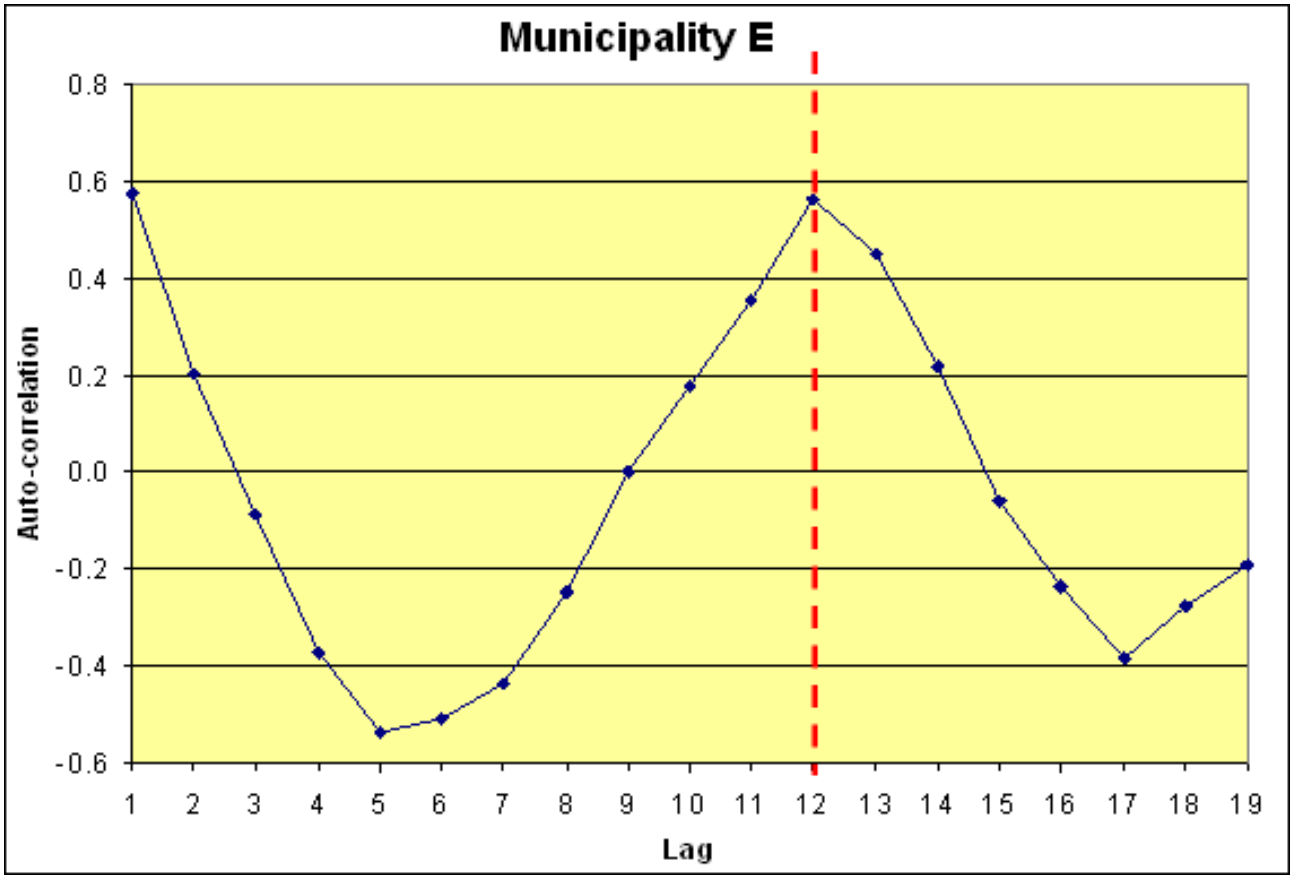
The significance level of the correlation at each lag L can be estimated using the normal standardised probability distribution z with:

$$Z_L = r_L \sqrt{(n - L)}$$

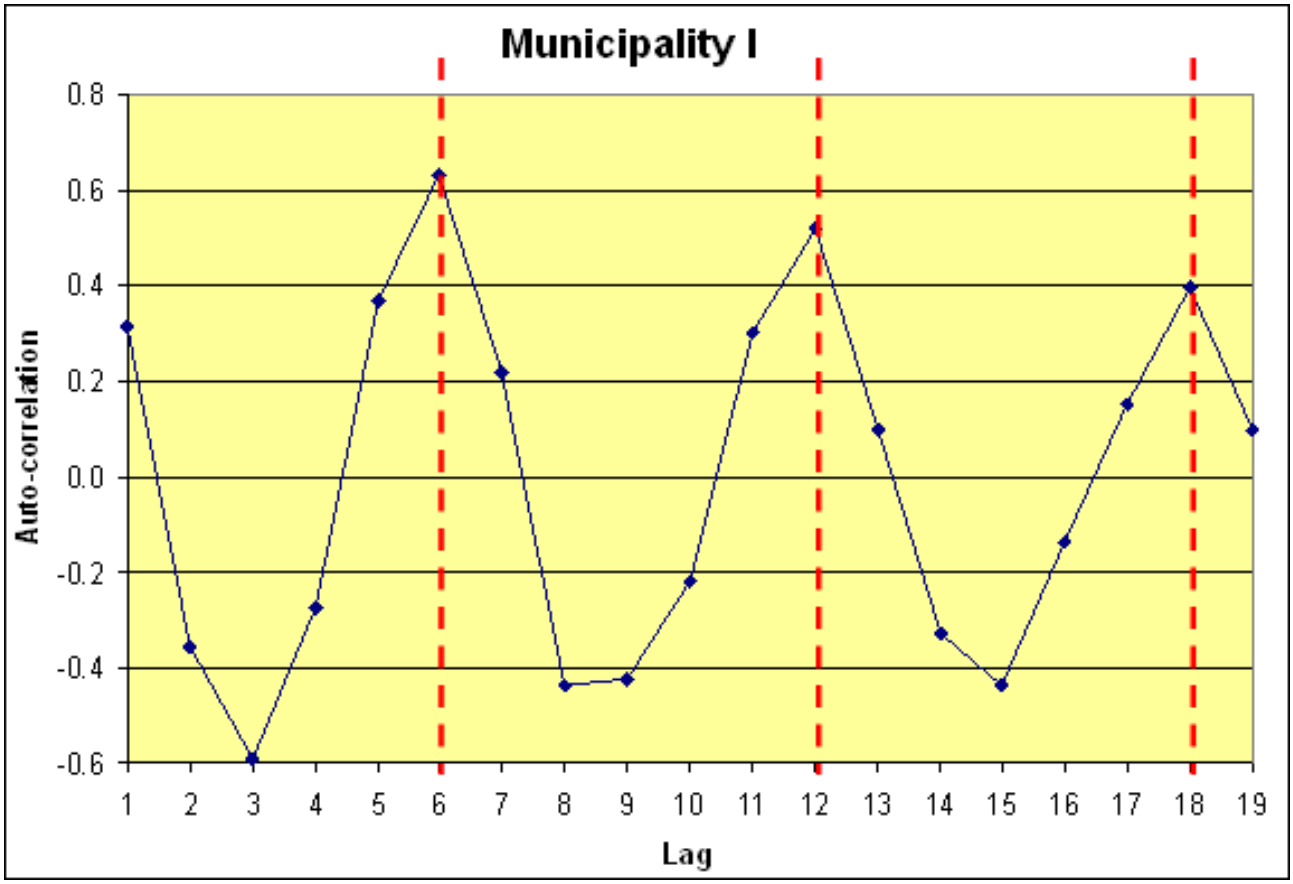
We can then plot the successive values in a diagramme of auto-correlation called *correlogramme*. Its interpretation will reveal the structure of the analysed time-series in terms of time influence decrease (rate of change) and presence of periodicities (cycles).

Let us now briefly illustrate this technique with the two time-series on the frequency of car accidents for municipalities E and I presented **here**. As each series is composed of 36 observations, the recommended maximum number of lags is about 9 (or n/4), however this number will be extended to 19, about the half of the series length, in order to better visualise possible cycles.

[16] A coefficient that expresses the correlation value between a time-series and itself at different time lags. Their scale of measurement must be at cardinal or ordinal level (see Lag, Cross-correlation)



Correlogrammes showing the periodicity of car accidents for municipality E



Correlogrammes showing the periodicity of car accidents for municipality I

Auto-association [17]:
At nominal level properties expressed by numerical values have no hierarchical meaning. Thus the only possible element of comparison between pairs of value is the "matching state". Within each compared pairs, property values are either identical (match) or different (mismatch). Thus an *index of similarity* [18] can be developed for measuring at each successive overlap position (match position, lag) the degree of similarity or association. Intuitively we can imagine this index as a ratio between the number of matching states and the number of comparisons:

[17] A procedure to measure the time dependancy of a phenomenon from a time-series compared to itself at different time lags. This technique is adapted for data measured at nominal level. (see Lag, Auto-correlation coefficient)

[18] An index expressing the degree of similarity or association between two time series or one by itself at different lag positions (see Auto-association)

$$a_L = m / n'$$

with:
m : number of matching states (matching pairs)
n' : number of comparisons (compared pairs)

Regardless of the number of compared pairs that change according to the lag value, this ratio varies between 1 and 0. A value of 1 indicates a perfect similarity or association as 0 indicates no association at all.
Similarly to the correlogramme, one can then plot the successive values in a diagramme of auto-association called *associatogramme*. Its interpretation will principally reveal the presence of periodicities (cycles) within the structure of the analysed time-series.
The significance level of the association at each lag L can be estimated using either a Chi-square test or an approximation of the binomial distribution (Davis 1986, p. 251). With a Chi-square test we compute a normalised difference between the number of observed matches in the sequence and the number of matches in a random sequence. This corresponds to the binomial probability of a given number of matches occurring when a random sequence is compared to itself. It is given by:

$$Pr = (\sum^{c}_{k=1} X^2_k) - n / (n^2 - n)$$

with:
c : number of categories (properties) in the observed sequence
n : length of the sequence (number of observations)

Once the probability of a match (Pr) for a random distribution is computed, one can deduce the probability of a mismatch Q as:

$$Q = 1 - Pr$$

We can now estimate the number of matches (E) and mismatches (E') occurring in a random sequence:

$$E = Pr * n'$$

$$E' = Q * n'$$

with:

n' : length of the compared sequence (number of compared pairs)

E : expected number of matches from a random sequence

E' : expected number of mismatches from a random sequence

It should be noted that the number of comparisons n' expresses the length of the effective compared sequence (overlapped segment) and therefore varies according to the offset (lag value).
We now have described all the components for the computation of the Chi-square value at each lag L:

Assuming this $\#^2$ test statistic has 1 degree of freedom, one can determine the significance of the association index value for each lag L. A Yates' correction factor can be applied to this statistic if the number of expected matches is small as with a reduced overlapping segment (Davis 1986, p. 250).
Let us now again briefly illustrate this technique with the time-series on the change in political majority during the period 1900-1990 for the municipality E presented in section **Methods for Time Series**. The series is composed of 10 observations, with 4 different properties (categories). Applying the previous rule, considered lag range from 0 to 3 for the computation of the index of similarity a_L . We have included lag 0 to illustrate the situation of comparing the time series with itself without any offset. The table below shows the pair comparison for the considered lag range.

Pol. Maj. E	Lag0	Lag1	Lag2	Lag3
1	1			
1	1	1		
2	2	1	1	
1	1	2	1	1
2	2	1	2	1
2	2	2	1	2
3	3	2	2	1
3	3	3	2	2
4	4	3	3	2
4	4	4	3	3
		4	4	3
			4	4
				4

Pairs compared for lags ranging from 0 to 3, discarded values are in grey.

Steps of the procedure are the following:

1. Computation of the index of similarity value a_L for lag 0 to lag 3

   a_L values can now being computed as the ratio of the number of matching pairs divided by the number of compared pairs. The detailed computation is shown below.

Lag#	n'	O	O'	a_L
0	10	10	0	1
1	9	4	5	0.44
2	8	2	6	0.25
3	7	2	5	0.29

2. Computation of the probability of matches in a random sequence

   We first have to calculate ($\#^c{}_{k=1}\ \#^2{}_k$) in the above formula:

k	X_k	X_k^2
1	3	9
2	3	9
3	2	4
4	2	4
	$\Sigma X_k^2 =$	26

   Finally Pr and Q have the following values for a sequence of 10 observations with 4 different properties:

   - Pr = (26 – 10) / (100 – 10) = 16 / 90 = 0.18
   - Q = 1 – 0.18 = 0.82

3. Computation of the Chi-square value for each lag L

   The table below details the computation of each Chi-square value:

Lag#	n'	O	O'	a_L	E	E'	O-E	O'-E'	$(O-E)^2$	$(O'-E')^2$	$(O-E)^2 / E$	$(O'-E')^2 / E'$	χ^2_L
0	10	10	0	1	1.8	8.2	8.2	-8.2	67.24	67.24	37.36	8.20	45.56
1	9	4	5	0.44	1.6	7.4	2.4	-2.4	5.76	5.76	3.60	0.78	4.38
2	8	2	6	0.25	1.4	6.6	0.6	-0.6	0.36	0.36	0.26	0.05	0.31
3	7	2	5	0.29	1.3	5.7	0.7	-0.7	0.49	0.49	0.38	0.09	0.46

4. Test of significance of Chi-square value for each lag L

   With a degree of freedom # = 1 and a confidence level of 95%, the critical Chi-square value is 3.84.
   We can then conclude that Chi-square values for lags 0 and 1 are significant as they are not for lags 2 and 3. In other words the auto-association is significantly different from a random sequence for lags 0 and 1 but not for lags 2 and 3. Its time dependency decreases rapidly but the short length of this illustrative series does not permit to observe the possible presence of change cycles. It is therefore not relevant to build up an associatogramme.

1.3. Multivariate time change analysis

What is the level of synchronisation between two phenomena or two features?
With multivariate analysis one can explore differences in evolution either between features or between phenomena for the same feature. When comparing the evolution of two features or two phenomena, behaviour differences correspond to *non similar evolution* trends or to a *time-lag*. When comparing time change, one should clearly identify the precise context of the analysis in order to select the most appropriate method:

- The richness of the time dimension: the time period is describe with simply *two limits* or with more details as a series of intervals constituting a *time series*.
- The level of measurement of the considered phenomena: nominal, ordinal or cardinal level. More simply as *qualitative* or *quantitative* data.
- The number of features or phenomena to be compared: with two, *pairwise comparison methods* (bivariate) can be chosen. With more than two, *multiple comparison methods* (multivariate) should be selected.

The following table lists 3 methods for multivariate analysis of the time dimension. One is used for comparing multiple features with several phenomena (variables) expressed at cardinal level but for only two time limits. The two remaining concern pairwise time series comparison adapted to qualitative or quantitative data.

	TIME DIMENSION	
	2 limits	Intervals
MULTIVARIATE	Change vector analysis (Quant)	Cross-association (Qual) Cross-correlation (Quant)

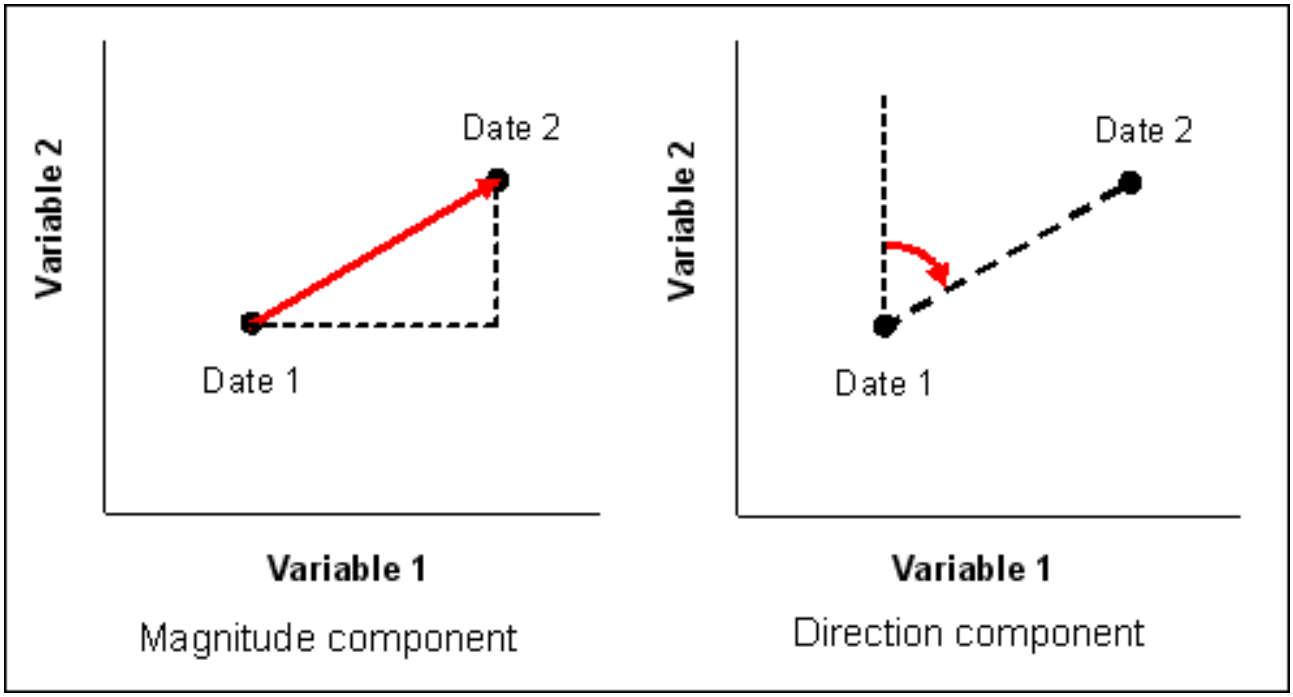
Examples of multivariate change analysis methods.

The principle of the change vector analysis method (CVA) will be briefly describe, as cross-association and cross-correlation methods will be illustrated with more details in the following sections.

1.3.1. Change vector analysis method (CVA)

The principle of *change vector analysis method (CVA)* [19] is to describe the change of individual feature across the different phenomena (variables) between two limits of time (two dates) as a vector within the variables space. Basically a vector can be described with a *magnitude* and a *direction* component. The magnitude component expresses the amount of change as the direction component informs about the type of change. The next figure illustrates the principle of change vector description within a two-dimensional variables space.

[19] The characteristics of a change in property value between two dates (moments) for 2 phenomena (variables) can be described as a vector expressing the strength (magnitude) of change as well as the direction of change with respect to the two variables



The two change vector components magnitude and direction describing the change of a feature between two time limits. Illustration for a two-dimensional variables space (adapted from Eastman, 2008, p.104).

Bivariate situation:

The analysis of time change in a bivariate situation corresponds to graphics in the last figure. This can be applied to simultaneously describe the change of properties of numerous features between two dates for two phenomena. A graphical representation as a scattergramme allows a comparison of changes between features to investigate. Change comparison is then based on three different characteristics:

- The *location* of the vector in this two dimensional space. It indicates the property values of each feature for the two variables and dates.
- The *magnitude* component expresses the amplitude of combined thematic change during the considered period of time. It indicates the individual dynamics of features.
- The *direction* component informs about the type of combined change between the two dates. It is measured as an angle clockwise from one variable axis, the variable 2 in the last figure.

Features can then be grouped into classes or categories of change behaviour according to their *magnitude* and *direction* values.

Multivariate situation:

When change analysis is concerned with more than two variables at a time, two strategies are available for using CVA method:

The number of original variables can be reduced to two components through a *principal component transformation*. This allows to return to a bivariate situation for the change vector analysis performed on these two first components. This approach is relevant when original variables are sufficiently correlated for producing a high degree of explained variations within the two first principal components. Otherwise this transformation leads to a significant loss of thematic information when undertaking the change vector analysis.

The bivariate change vector method can be extended to a multivariate situation. One can imagine a variable space with not only two dimensions, but *n* dimensions. The multivariate change behaviour of each feature can still be described as a vector with a single magnitude index, but with several direction indices, in fact *n-1*. The magnitude component can still be interpreted as the individual dynamics of features and direction components still express in a more complex manner the type of multivariate change occurring during this period of time.

1.3.2. Cross-correlation

We know that the correlation techniques express the amount of synchronised change between two phenomena or variables. When applied to the time dimension one can compare property changes of an individual feature for two different variables during the same period of time. Similarly it is possible to compare the behaviour change of two different features for the same variable. This is obtained by comparing two time-series. For both situations one expects to discover a significant similarity between the two variables or the two features during the considered period of time. The hypothesis of a significant relationship between features or variables can be validated only when change is synchronised within the considered period of time. Let us suppose that two phenomena are strongly correlated, but with a time-lag between them or that two features are influenced by a same factor but with a different time response and speed. Thus a simple correlation procedure that compares the two time-series values on a date-by-date basis will indicate a very low degree of correlation. We have seen in section **Time dependency** a technique called auto-correlation that is capable of comparing a time-series with itself with different time-lag values. One can then imagine to apply this principle for the comparison of two variables or two features shifted forward or backward. This can be performed with a technique called *cross-correlation* [20]. The context of use is slightly different and more complex than the one of the auto-correlation:

- The *length* of the two compared time-series might be different.
- To fully investigate possible shifts in time between the two series one should consider not only positive time-lags but also those ones negative. We will then prefer the term of *match positions* rather than lags to describe the successive comparisons.
- The equation for cross-correlation differs slightly from the auto-correlation index, but still refers to the Pearson linear correlation coefficient. If the two series are called Y_1 and Y_2 and the number of compared pairs (overlapped positions) between the two chains at the match position *p* is designated as *n'*, then the equation can be written as follow:

$$r_p = COV'(Y_1, Y_2) / s'_{y1} * s'_{y2}$$

with:

COV' : covariance of overlapped segments of the two chains (with n' pairs)

s'_{y1} , s'_{y2} : standard deviations of overlapped segments of the two chains

r_p : cross-correlation coefficient at match position p

[20] A procedure to measure the correlation value between two time-series at different time lags. Their scale of measurement must be at cardinal or ordinal level (see Lag, Cross-association)

- The significance of the cross-correlation coefficient at each match position m can be evaluated by an approximate test derived from a test developed for the correlation coefficient:

$$t_p = r_p \sqrt{((n' - 2) / (1 - r^2_p))}$$

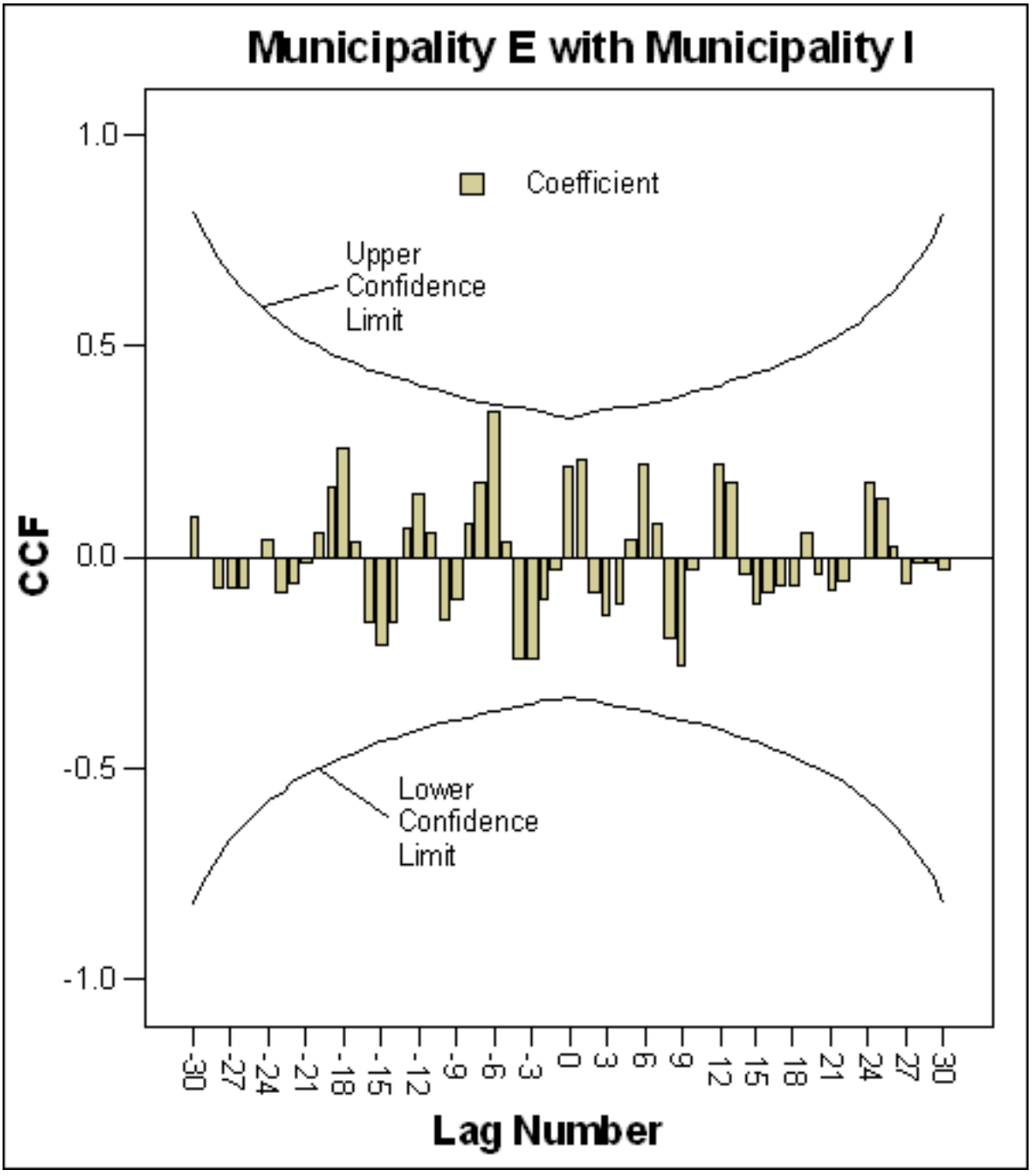
with:

n' : number of overlapped positions between the two chains (compared pairs)

The degrees of freedom # equals n'-2 and the null hypothesis states that the cross-correlation is not significantly different from zero.

Let us return to the evolution of the number of car accidents in municipalities E and I during the period of 26 months. This **figure** illustrates the regularity of car accidents peak every 6 months for municipality I and every 12 months for municipality E. When comparing their evolution pattern with cross-correlation technique, one can expect to observe the following:

- There is no time-shift between peaks for the two municipalities as they match every month of July.
- Around February the difference between the two municipalities is the strongest.
- Cycle length is about 6 and 12 months for municipality I and E respectively.



Cross-correlation coefficients (CCF) computed for the two time-series on the number of car accidents in municipalities E and I. Match positions are in the range of –30 to +30.

EXERCISE

From the last figure and with the help of the distribution of monthly car accidents in the municipalities E and I during the period 1900-1990 (**figure**), try to confirm the above three observations made about the two time-series.

1.3.3. Cross-association

The *cross-association* [21] is the alternative technique for the pairwise comparison of change pattern between either two different features or two variables of the same feature, when the level of measurement is nominal (qualitative data). The comparison process between the two time-series is similar to the cross-correlation technique, but the measure of correspondence (degree of correspondence) at each match position is the index of similarity. It is the same ratio as the *index of similarity* a_L computed for the auto-association, but this time it is called a_p as match position can be negative or positive:

$$a_p = m / n'$$

with:
m : number of matching states (matching pairs)
n' : number of comparisons (compared pairs)

Again the ratio range between 0 and 1, expressing the strength of similarity between the two segments compared at match position m. Successive values can be plotted in an *associatogramme* to identify match positions with a high degree of similarity and the to interpret the related shifts in time.
The significance level of the association at each match position m can again be estimated using either a Chi-square test or an approximation of the binomial distribution. Just like for the auto-association case, with a Chi-square test we compute a normalised difference between the number of observed matches in the segment of sequences and the number of matches in a random sequence. But in this context it corresponds to the binomial probability of a given number of matches occurring when two random sequences are compared. It is given by:

$$Pr = (\sum^{c}_{k=1} X_{1k} * X_{2k}) / (n_1 * n_2)$$

with:
c : number of categories (properties) in the observed sequences
n_1 : total length of the sequence 1 (number of observations)
n_2 : total length of the sequence 2 (number of observations)

Then the following steps are identical to those applied for auto-association index evaluation. Once the probability of a match (Pr) for a random distribution is computed, one can deduce the probability of a mismatch Q as:

$$Q = 1 - Pr$$

We can now estimate the number of matches (E) and mismatches (E') occurring in a random sequence:

[21] A procedure to measure the correlation value between two time-series at different time lags. Their scale of measurement must be at nominal level (see Lag, Cross-correlation)

$$E = Pr * n'$$

$$E' = Q * n'$$

with:

n' : length of the compared sequence (number of compared pairs)

E : expected number of matches from a random sequence

E' : expected number of mismatches from a random sequence

It should be noted that the number of comparisons n' expresses the length of the effective compared sequence (overlapped segment) and therefore varies according to the match position.

We now have described all the components for the computation of the Chi-square value at each match position p:

$$\chi^2_p = ((O - E)^2 / E) + ((O' - E')^2 / E')$$

with:

O : observed number of matches

O' : observed number of mismatches

E : expected number of matches from a random sequence

E' : expected number of mismatches from a random sequence

Assuming this #^{2} test statistic has 1 degree of freedom, one can determine the significance of the association index value for each match position p.

Let us illustrate this procedure of cross-association with the comparison of political majority change between municipalities A and E during the period 1900-1990 (**Table**). The two series being similar in size –same period- but describing two different features, comparison should be applied at both negative and positive shifted positions as illustrated in the following table.

	Political Majority Municipality A														
Pol. Maj. E	p-7	p-6	p-5	p-4	p-3	p-2	p-1	p 0	p+1	p+2	p+3	p+4	p+5	p+6	p+7
	2														
	1	2													
	1	1	2												
	2	1	1	2											
	2	2	1	1	2										
	4	2	2	1	1	2									
	3	4	2	2	1	1	2								
1	**3**	**3**	**4**	**2**	**2**	**1**	**1**	**2**							
1	**3**	**3**	**3**	**4**	**2**	**2**	**1**	**1**	**2**						
2	**1**	**3**	**3**	**3**	**4**	**2**	**2**	**1**	**1**	**2**					
1		**1**	**3**	**3**	**3**	**4**	**2**	**2**	**1**	**1**	**2**				
2			**1**	**3**	**3**	**3**	**4**	**2**	**2**	**1**	**1**	**2**			
2				**1**	**3**	**3**	**3**	**4**	**2**	**2**	**1**	**1**	**2**		
3					**1**	**3**	**3**	**3**	**4**	**2**	**2**	**1**	**1**	**2**	
3						**1**	**3**	**3**	**3**	**4**	**2**	**2**	**1**	**1**	**2**
4							**1**	**3**	**3**	**3**	**4**	**2**	**2**	**1**	**1**
4								**1**	**3**	**3**	**3**	**4**	**2**	**2**	**1**
									1	3	3	3	4	2	2
										1	3	3	3	4	2
											1	3	3	3	4
												1	3	3	3
													1	3	3
														1	3
															1

compared segments

Comparison of political majority change between municipalities A and E. Pairs compared for match positions ranging from -7 to +7, compared segments are in bold.

Steps of the procedure are very similar to the auto-association technique, they are the following:

1. Computation of the **index of similarity** a_p for match position -7 to +7
   a_p values can now being computed as the ratio of the number of matching pairs divided by the number of compared pairs. The detailed computation is shown below.

p#	n'	O	O'	a_p
-7	3	0	3	0
-6	4	1	3	0.25
-5	5	0	5	0
-4	6	0	6	0
-3	7	0	7	0
-2	8	3	5	0.38
-1	9	5	4	0.56
0	10	4	6	0.4
1	9	4	5	0.44
2	8	3	5	0.38
3	7	1	6	0.14
4	6	2	4	0.33
5	5	1	4	0.2
6	4		4	0
7	3		3	0

2. Computation of the **probability of matches** in a random sequence

   We first have to calculate $\#^{c}_{k=1} (X_{1k}*X_{2k})$ in the above formula:

k	X_{1k}	X_{2k}	$X_{1k}X_{2k}$
1	3	3	9
2	3	3	9
3	2	3	6
4	2	1	2
$\Sigma X_{1k}X_{2k} =$			26

   Finally Pr and Q have the following values for a sequence of 10 observations with 4 different properties:

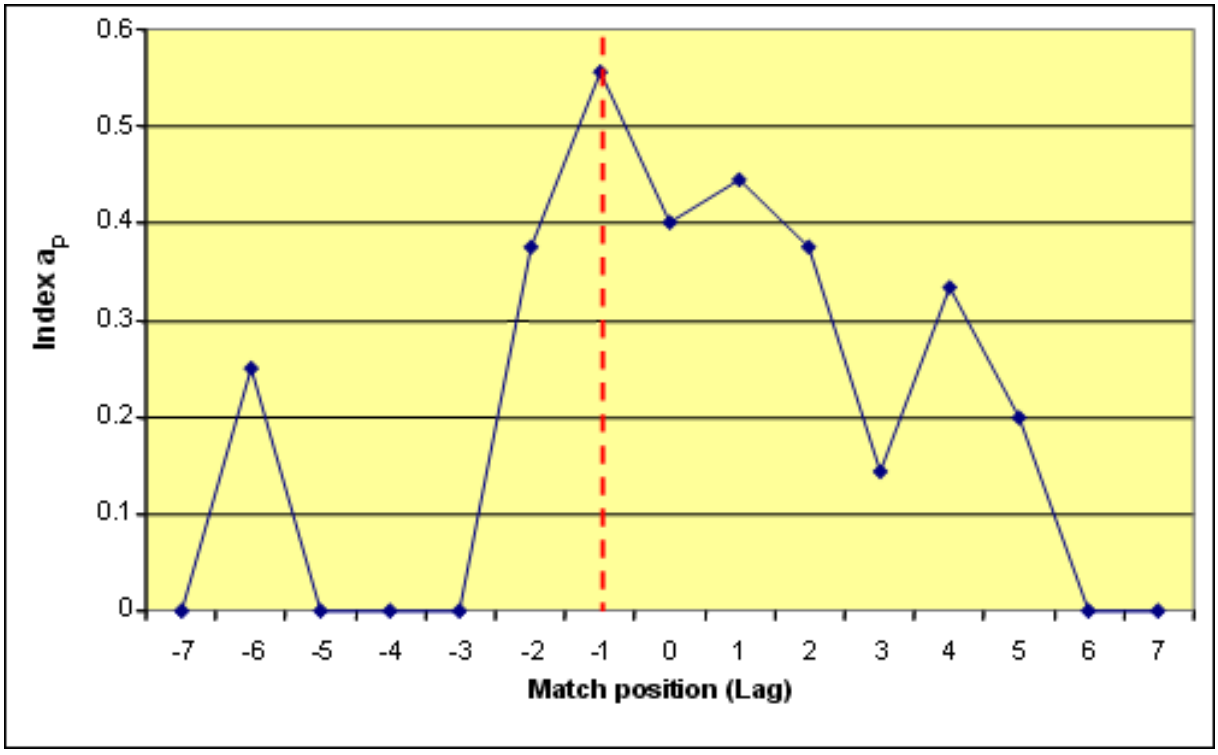
   - Pr = 26 / (10 * 10) = 26 / 100 = 0.26
   - Q = 1 – 0.26 = 0.74

3. Computation of the **Chi-square** value for each match position p

   The table below details the computation of each Chi-square value:

p#	n'	O	O'	a_p	E	E'	O-E	O'-E'	$(O-E)^2$	$(O'-E')^2$	$(O-E)^2$ / E	$(O'-E')^2$ / E'	χ^2_p
-7	3	0	3	0	0.78	2.22	-0.78	0.78	0.6084	0.6084	0.78	0.27	1.05
-6	4	1	3	0.25	1.04	2.96	-0.04	0.04	0.0016	0.0016	0.00	0.00	0.00
-5	5	0	5	0	1.3	3.7	-1.3	1.3	1.69	1.69	1.30	0.46	1.76
-4	6	0	6	0	1.56	4.44	-1.56	1.56	2.4336	2.4336	1.56	0.55	2.11
-3	7	0	7	0	1.82	5.18	-1.82	1.82	3.3124	3.3124	1.82	0.64	2.46
-2	8	3	5	0.38	2.08	5.92	0.92	-0.92	0.8464	0.8464	0.41	0.14	0.55
-1	9	5	4	0.56	2.34	6.66	2.66	-2.66	7.0756	7.0756	3.02	1.06	4.09
0	10	4	6	0.4	2.6	7.4	1.4	-1.4	1.96	1.96	0.75	0.26	1.02
1	9	4	5	0.44	2.34	6.66	1.66	-1.66	2.7556	2.7556	1.18	0.41	1.59
2	8	3	5	0.38	2.08	5.92	0.92	-0.92	0.8464	0.8464	0.41	0.14	0.55
3	7	1	6	0.14	1.82	5.18	-0.82	0.82	0.6724	0.6724	0.37	0.13	0.50
4	6	2	4	0.33	1.56	4.44	0.44	-0.44	0.1936	0.1936	0.12	0.04	0.17
5	5	1	4	0.2	1.3	3.7	-0.3	0.3	0.09	0.09	0.07	0.02	0.09
6	4		4	0	1.04	2.96	-1.04	1.04	1.0816	1.0816	1.04	0.37	1.41
7	3		3	0	0.78	2.22	-0.78	0.78	0.6084	0.6084	0.78	0.27	1.05

4. Test of **significance** of Chi-square value for each match position m With a degree of freedom v = 1 and a confidence level of 95%, the critical Chi-square value is 3.84.
We can then conclude that only the Chi-square value at match position -1 is significant. In other words the cross-association is significantly different from a random sequence for a negative time-shift of 1 between the two sequences. This can be observed when comparing the column “Pol. Maj. E” with the column “m-1” of “Political Majority Municipality A” in the Table 2.32. The cross-association coefficients ap for each of considered match positions can then be plotted as an associatogramme (Next figure). It shows that largest similarities between the two series occur between match positions –2 and +2.



Associatogramme showing the successive cross-association coefficient values for match positions in the range of –7 to +7. The two sequences are most similar with a time-shift corresponding to one year.

EXERCISE

From table expressing political majority change between municipalities A and E during the period 1900-1990 (**Table**) try to visually estimate:

- The two municipalities with the highest similarity at match position 0.
- Two series with a high degree of similarity but with a time shift either negative or positive.

1.4. Summary

The objective of a thematic change analysis is to explore, understand and/or to forecast changes in the properties of spatial features. Since only the thematic and the time dimensions are concerned, many “standard” methods and techniques are offered to analyse such changes. Therefore, the main challenge is to select the most appropriate methodology that satisfies our objectives in a specific informational context. One can describe either the overall pattern of change within a set of spatial features throughout a period of time, or the analysis can concentrate on their individual change characteristics.

This Unit proposes a framework to organise potential methods with respect to the objectives and to the information context of the change analysis. It is structured into three sections: the production of change indices, the description of the behaviour of time series, and multivariate time change analysis. Such approaches are mainly concerned with the thematic changes of the properties of spatial features, both at univariate and multivariate levels.

1.5. Recommended Reading

- **Abler R., Adams J., Gould P.**, 1972. *Spatial Organization, The Geographer's View of the World*. USA: Prentice Hall.
- **Caloz R., Collet C.**, 2011. *Analyse spatiale de l'information géographique*. Lausanne, Switzerland: PPUR.
- **Davis, J.C.**, 1986. *Statistics and data analysis in geology*. New York: John Wiley &Sons.
- **Eastman, R.**, 2008. *IDRISI - Taiga GIS and Image processing software, Reference Manual*. Worcester, Clark Labs, Clark University, USA: Worcester.

1.6. Glossary

Allometric function:

A regression function that describes the growth rate of a part with respect to the growth of the entire organism (see Allometry)

Allometry:

Allometry is a concept developed in biology. “Allometry: the relative growth of a part in relation to an entire organism or to a standard” (Merriam-Webster)

Auto-association:

A procedure to measure the time dependancy of a phenomenon from a time-series compared to itself at different time lags. This technique is adapted for data measured at nominal level. (see Lag, Auto-correlation coefficient)

Auto-correlation (temporal):

A procedure to measure the time dependancy of a phenomenon from a time-series compared to itself at different time lags. (see Lag, Auto-correlation coefficient)

Auto-correlation coefficient:

A coefficient that expresses the correlation value between a time-series and itself at different time lags. Their scale of measurement must be at cardinal or ordinal level (see Lag, Cross-correlation)

Change index (global):

A change index is an indicator derived from multitemporal measurements. It expresses the amount of change within a period of time. It can describe the change behavior of a set of features (global) or of individual features. It can result from a difference, a ratio, ...

Change vector analysis (CVA):

The characteristics of a change in property value between two dates (moments) for 2 phenomena (variables) can be described as a vector expressing the strength (magnitude) of change as well as the direction of change with respect to the two variables

Correlogramme:

A graphical representation of a succession of correlation values varying throughout time or space (see Lag, Auto/Cross-correlation)

Cross-association:

A procedure to measure the correlation value between two time-series at different time lags. Their scale of measurement must be at nominal level (see Lag, Cross-correlation)

Cross-correlation:

A procedure to measure the correlation value between two time-series at different time lags. Their scale of measurement must be at cardinal or ordinal level (see Lag, Cross-association)

Lag:

In time series analysis, variables can be compared synchronously or with a defined time lag. As time series are generally made of regularly distributed intervals of time, this **asynchronous comparison** corresponds to one or several lag steps

Linear regression function:

A regression function that relates a dependant variable Y with one or several independant variables Xi in a linear manner. A first degree polynomial function is a linear function (see Polynomial regression function)

Markov chain (analysis):

A technique to estimate the probability of occurrence from any original state to any final state after a specific sequence of n time steps. It makes use of transition matrices

Periodicity:

A succession of properties that occurs regularly throughout time. For example daily temperatures or seasonal unemployement

Principal component analysis (PCA):

A procedure that transforms an original set of variables into a set of Principal components. This transformation removes the original correlation between variables (information redundancy) and structure the overall variability into ordered components (the first component carrying more variability than the second, and so on)

Runs test:

A runs test aims to compare an observed time series with a random sequence of states

Similarity index:

An index expressing the degree of similarity or association between two time series or one by itself at different lag positions (see Auto-association)

Thematic properties:

Values attached to observations expressing their property for each considered phenomenon (variable)

Time series:

A sequence of measurements ordered according to Time (moments of time). It describes the change of properties of a single observation throughout time

Transition matrix:

A general term to identify any matrix that expresses a change of properties (states) between two moments

Transition probability matrix:

A probability matrix that expresses a transition from one state to another

Transition proportion matrix:

A matrix that expresses the tendency of one state to follow another

Trend surface (analysis):

A regression function modelling the property values (Z) based on their location (X,Y) in space: $Z = f(X,Y)$

1.7. Bibliography

- **Anonymous**. *Allometry* [online]. Available from: www.Merriam-Webster/dictionary/allometry.
- **Abler R., Adams J., Gould P.**, 1972. *Spatial Organization, The Geographer's View of the World*. USA: Prentice Hall.
- **Caloz R., Collet C.**, 2011. *Analyse spatiale de l'information géographique*. Lausanne, Switzerland: PPUR.
- **Davis, J.C.**, 1986. *Statistics and data analysis in geology*. New York: John Wiley &Sons.
- **Eastman, R.**, 2008. *IDRISI - Taiga GIS and Image processing software, Reference Manual*. Worcester, Clark Labs, Clark University, USA: Worcester.