ASSESSING INHOMOGENEOUS INDICATOR-BASED TYPOLOGIES THROUGH THE REVERSE CLUSTERING APPROACH

Jan W. Owsiński, Jarosław Stańczak, Sławomir Zadrożny and Janusz Kacprzyk

Systems Research Institute, Polish Academy of Sciences

owsinski@ibspan.waw.pl, stanczak@ibspan.waw.pl, zadrozny@ibspan.waw.pl, Janusz.Kacprzyk@ibspan.waw.pl

12 POW Neuchatel October 2018

The content

- Reverse clustering – what is it? how is it done?
- The problem at hand: the typology of municipalities for planning purposes / verifying the typology in data analysis (clustering) context – the rationale
- The data
- The results and some conclusions
- A broader picture and potential application domain(s)

The general problem and the reverse clustering approach (1)

We are given a partition ***P_A*** (composed of p_A subsets) of a given set X of objects, x_i, indexed i, $i = \{1,\ldots,n\}$, described by vectors of values x_i.

Reverse clustering: find the parameters of the clustering procedure such that the procedure, defined by these parameters, leads to a partition ***P_B*** of the given set of objects that is possibly the closest (most similar) to P_A (minimise the „distance" between P_A and P_B).

The parameters thereby optimised include:

1. choice of the algorithm;

2. key parameter(s) of the algorithm;

3. weights or choice of variables; and

4. definition of distance (e.g. the Minkowski exponent).

The general problem and the reverse clustering approach (2)

Similarity of P_A and P_B is measured with Rand / adjusted Rand index, possibly with a regularising component.

The clustering algorithms accounted for:

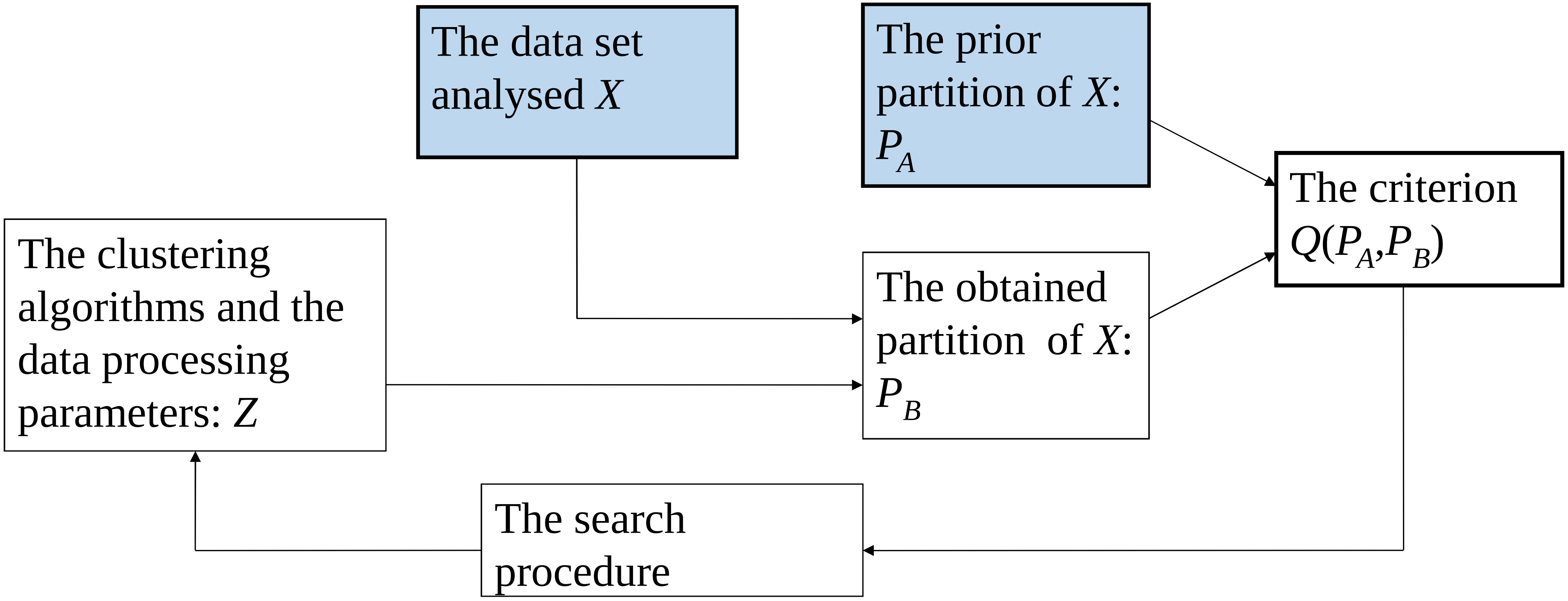
-- k-means / k-medians (parameterised with the numer of clusters);

-- DBSCAN (parameterised with the numer of neighbours and maximum distance); and

-- general progressive merger (parameterised with Lance-Williams formula).

The vector of „best" parameters is sought with evolutionary algorithms: own evolutionary algorithm – two-level adaptation (operators & individuals)

The procedure



The data set analysed X

The prior partition of X: P_A

The criterion $Q(P_A,P_B)$

The clustering algorithms and the data processing parameters: Z

The obtained partition of X: P_B

The search procedure

The issues and the understanding

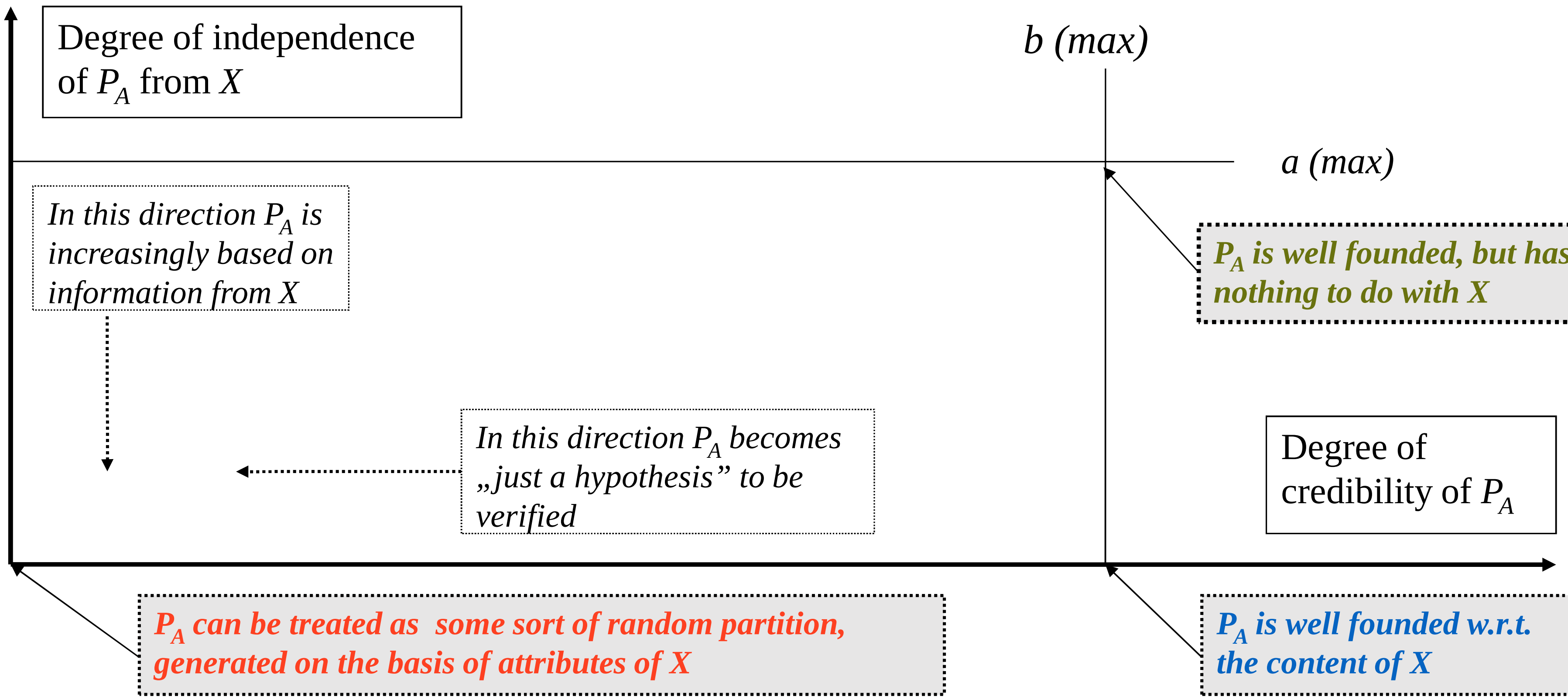
Partition P_A can be considered a „model”, of various potential characteristics, which ***we wish to reconstruct within a definite methodological domain (here: clustering)***.

The problem has two aspects: the *technical* and the *substantive* ones.

- Technical: the perfection of the search for the approximation of P_A
- Substantive: the use of the approximating parameters found, e.g. for clustering much bigger data sets, for drawing conclusions from the differences between P_A and the approximating partition, for finding special subgroups (e.g. those defining the difference), etc.

What is the „inner sense” of the procedure / approach / problem?

The understanding (1)



Degree of independence of P_A from X

b (max)

a (max)

In this direction P_A is increasingly based on information from X

P_A is well founded, but has nothing to do with X

In this direction P_A becomes „just a hypothesis" to be verified

Degree of credibility of P_A

P_A can be treated as some sort of random partition, generated on the basis of attributes of X

P_A is well founded w.r.t. the content of X

The understanding (2)

The original purpose (now just one of many…):

To provide the mechanism for categorising the objects in other, generally / roughly similar, but yet different data sets. Especially <u>much bigger</u> data sets.

In particular: not the one-by-one classification of the incoming objects.

The fundamental distinction

The question:

Is this not (simply) another method of determining classifier(s)?
Why not try out known methods of classification?

The answer: No. Why? Because:

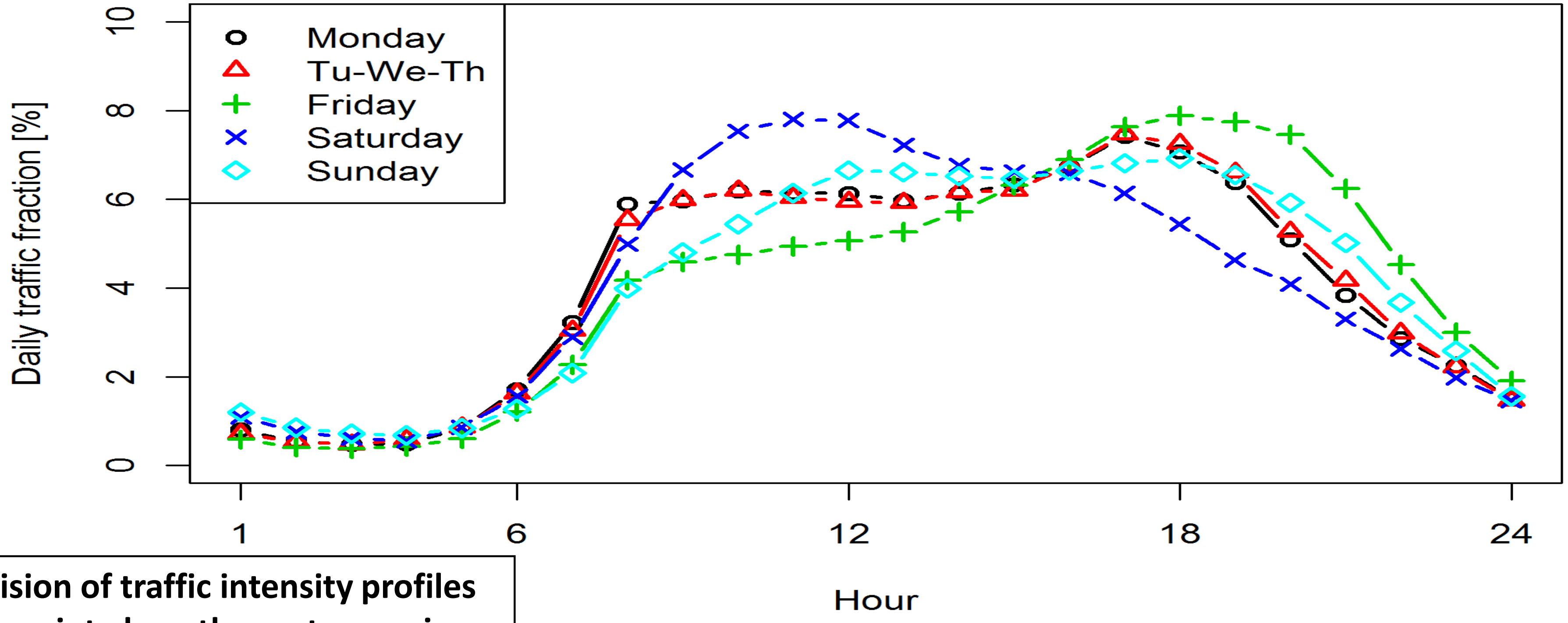
1. **We aim at classifying „at once" relatively large data sets (the question is not „*where a given observation belongs*", but „*how to divide a given data set*").**
2. **We wish the „classifier" to allow for more flexibility and provision of additional information (e.g. different number of clusters than in the initial partition, outliers,…).**

The search procedure

In view of the cumbersome „landscape” of the solution space, highly nonlinear, (dis)continuous-discrete etc., the search methodology of choice is <u>evolutionary optimisation</u>.

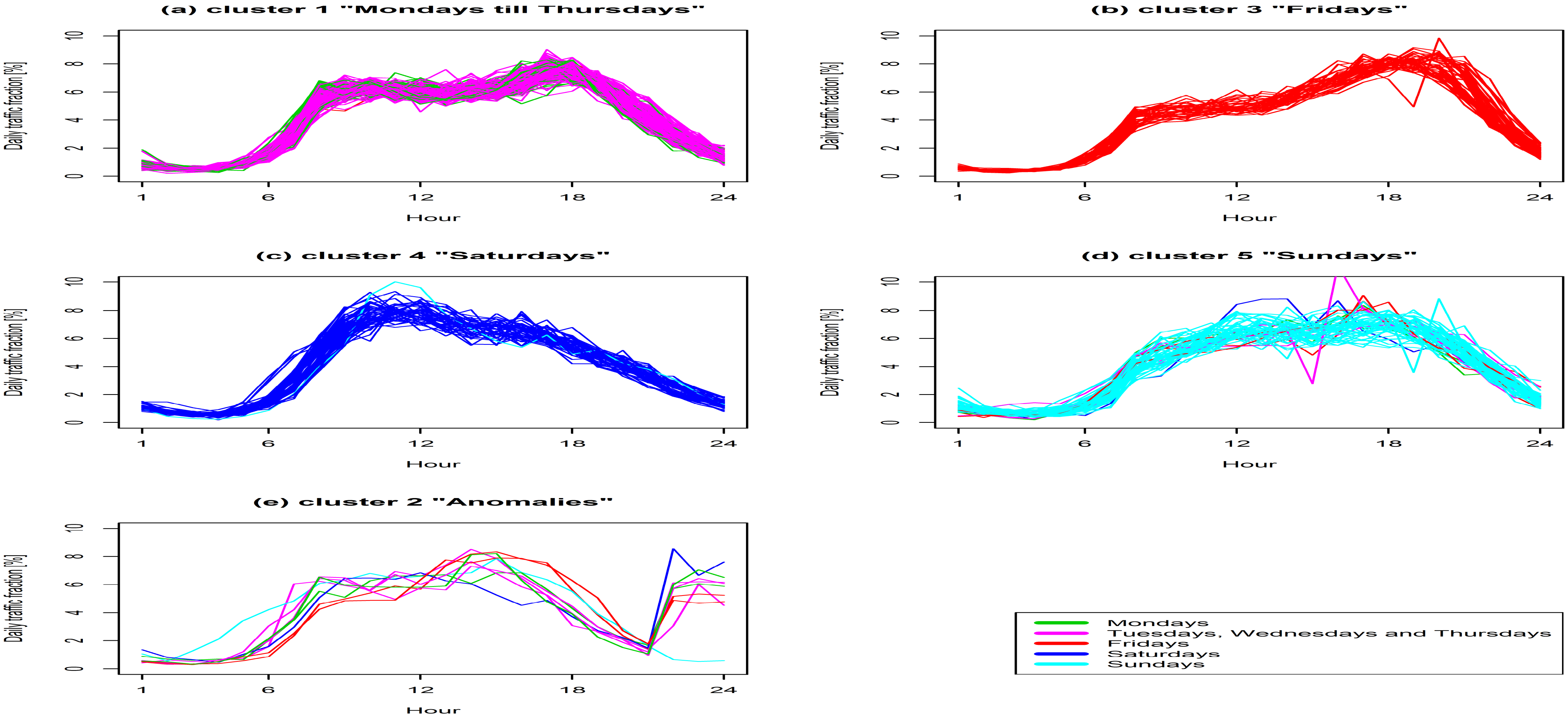
The algorithm applied, of own development, is a <u>two-level</u> one: the usual population evolution level + operator assessment and choice level (each individual descendance line is [also] characterised by operator assessment coefficients, helping in selecting operators at each step).

An example (1)



Division of traffic intensity profiles at a point along the motorway in Germany according to the days of the week

An example (2)


(a) cluster 1 "Mondays till Thursdays"
(b) cluster 3 "Fridays"
(c) cluster 4 "Saturdays"
(d) cluster 5 "Sundays"
(e) cluster 2 "Anomalies"
Daily traffic fraction [%]
0
2
4
6
8
10
1
6
12
18
24
Hour
Mondays
Tuesdays, Wednesdays and Thursdays
Fridays
Saturdays
Sundays

An example (3)

Results for traffic data for the entire vector of parameters, obtained with the use of hierarchical aggregation (Rand index = 0.850, adjusted Rand = 0.654).

Prior partition (*P_A*):	Clusters obtained (*P_B*):				
	1 (Mon-Tu-Wed-Th)	2 (outliers)	3 (Friday)	4 (Saturday)	5 (Sunday)
Friday	1	2	**42**	0	3
Monday	**45**	2	0	0	2
Saturday	0	1	0	**46**	1
Sunday	0	1	0	1	**47**
Tu-We-Th	**140**	3	0	0	4

The problem at hand (1)

Typology of some 2 500 Polish municipalities for definite planning purposes

1. A typology was developed by the specialists from the Institute of Geography and Spatial Organization of the Polish Academy of Sciences for planning purposes
2. The typology was based on (a) a spectrum of variables, (b) some administrative criteria, (c) some functional criteria (e.g. transport or other special sectors of economy), forming a definite, **branching procedure**
3. The exercise consists in the attempt to reconstruct this typology on the basis of a set of apparently tangible variables, which could then serve to possibly (i) modify the original types, (ii) establish alternative, more „objective" typology, and (iii) identify the criteria used in the original typology that are most „twisting" the counterpart quasi-objective one, obtained with a data analysis procedure
4. The exercise was carried out with the reverse clustering approach, using the evolutionary algorithm.

The problem at hand (2)

Typology of some 2 500 Polish municipalities for definite planning purposes

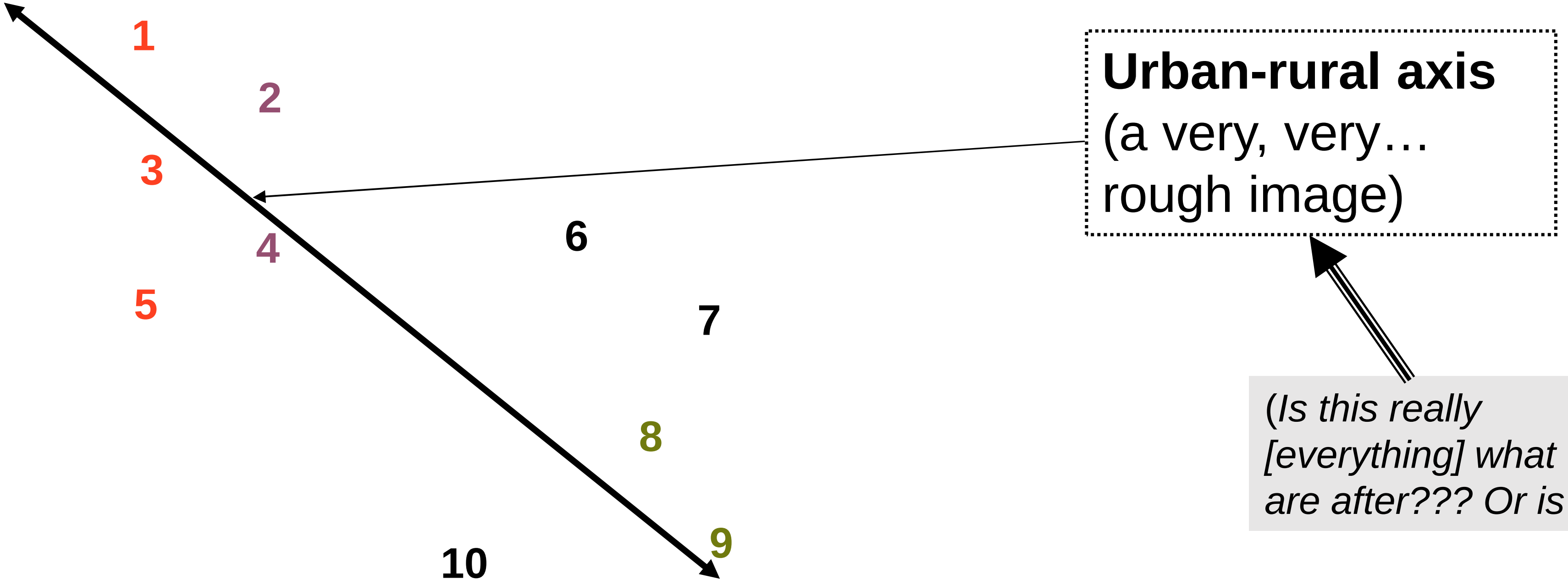
Types distinguished in the original typology:

1	functional urban areas (fua’s) of voivodship capitals
2	external zones of fua's of voivodship capitals
3	functional urban areas of subregional centres
4	external zones of fua's of subregional centres
5	multifunctional urban centres (other)
6	communes with developed transport functions
7	communes with other developed non-farming functions (tourism and large-scale functions, including mining)
8	communes with intensive farming functions
9	communes with moderate farming functions
10	extensively developed communes (with forests or nature protection areas)

The problem at hand (3)

Typology of some 2 500 Polish municipalities for definite planning purposes

Types distinguished in the original typology (a sort of mapping):



Urban-rural axis (a very, very… rough image)

(Is this really [everything] what we are after??? Or is it???)

The problem at hand (4)

The variables used to carry out the reverse clustering (characteristics of municipalities):

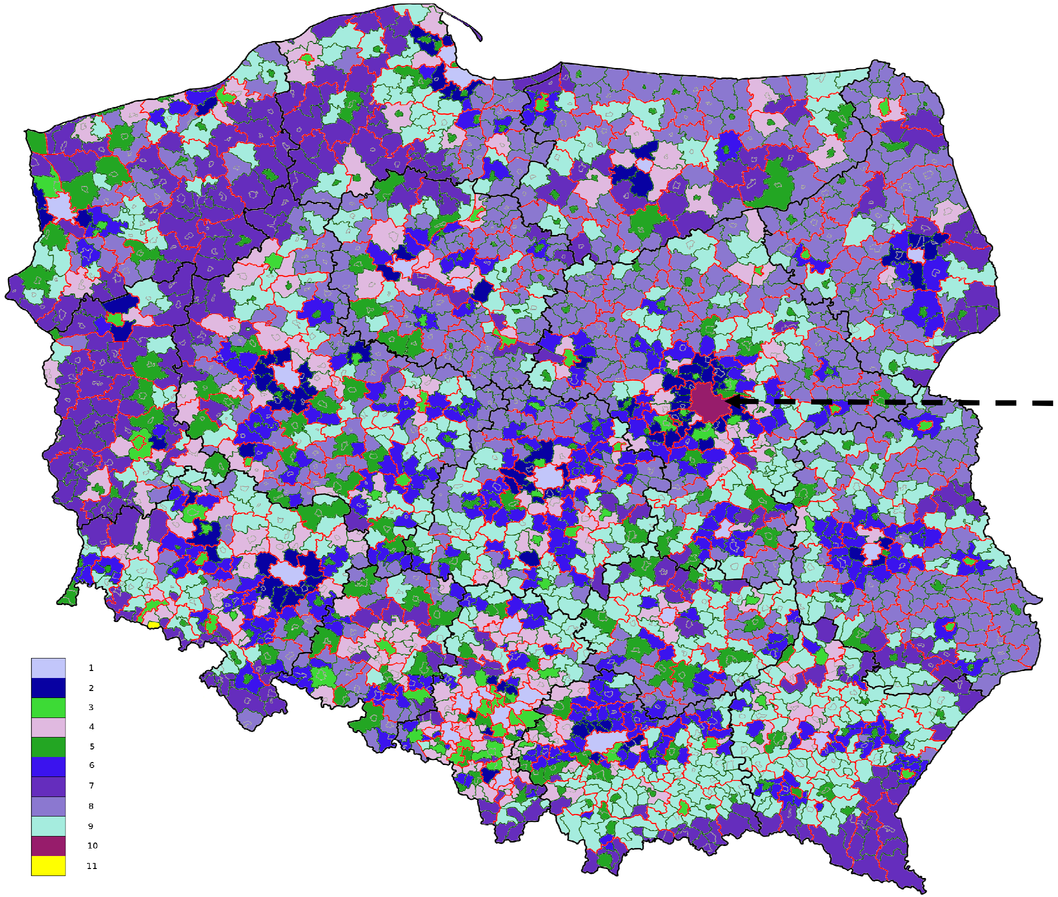
1. **Population number**
2. **Overbuilt area**
3. **Share of transport related areas**
4. **Population density**
5. **Share of agricultural land**
6. **Share of overbuilt areas**
7. **Share of forest areas**
8. **Share of population over 60 years of age**
9. **Share of population below 20 years of age**
10. **Birthrate for the last 3 years**
11. **Migration balance for the last 3 years**
12. **Average farm acreage indicator**
13. **Registered employment indicator**
14. **Registered businesses per 1 000 inhabitants**
15. **Employment-based average business magnitude indicator**
16. **Share of businesses from manufacturing and construction**
17. **Number of pupils per 1 000 inhabitants**
18. **Number of students of over-primary schools per 1 000 inhabitants**
19. **Own revenues of municipality per inhabitant**
20. **Share of revenues from personal income tax in own communal revenues**
21. **Share of social care expenses in total communal budget**

The problem at hand (5)

Typology of some 2 500 Polish municipalities for definite planning purposes

Can we count on the re-establishment of the original typology?

Why?



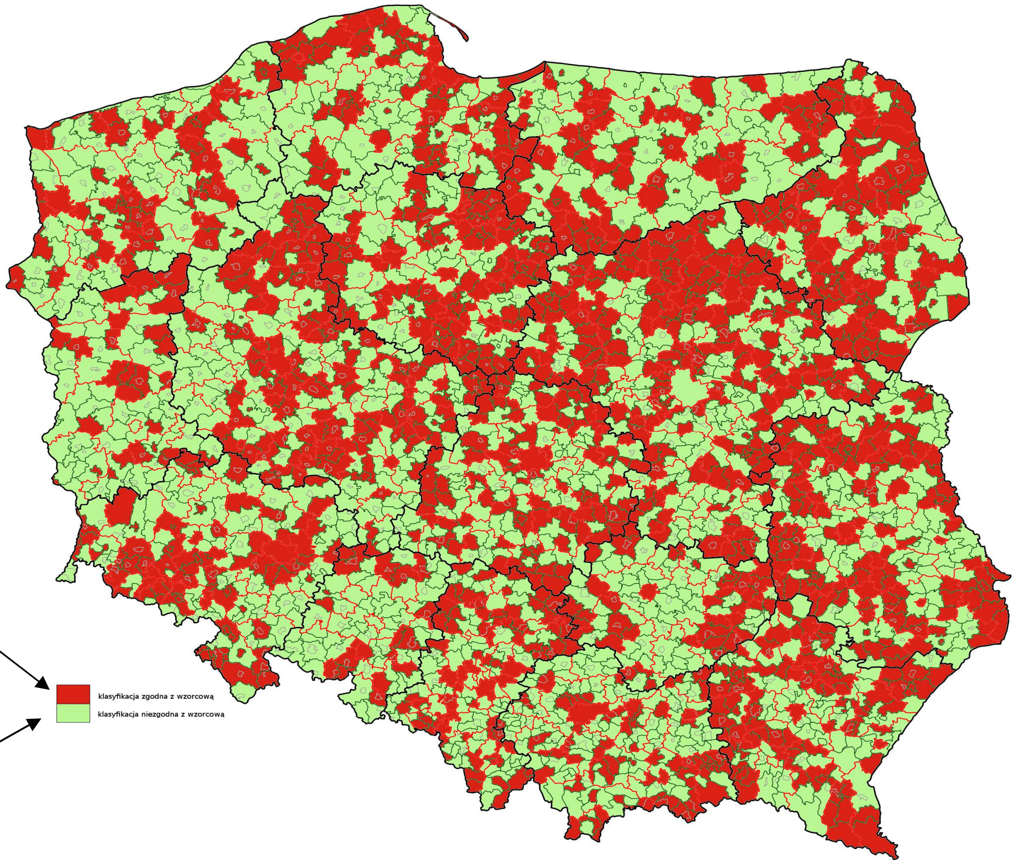
One of the solutions obtained: k-means, 11 clusters

A new 10th cluster…

The problem at hand (6)

Typology of some 2 500 Polish municipalities for definite planning purposes

Can we count on the re-establishment of the original typology? Why?


klasyfikacja zgodna z wzorcową
klasyfikacja niezgodna z wzorcową

communes classified **conform to the original typology**

communes classified **differently than in the original typology**

One of the solutions obtained: k-means, 11 clusters

The problem at hand (7)

Typology of some 2 500 Polish municipalities for definite planning purposes

The confusion matrix between the original and [one of] the „best” obtained partitions:

Obtained categories (clusters) of communes:

Given categories of municipalities and their interpretations:

		1	2	3	4	5	6	7	8	9	10	11	error sum	error share	totals
1	functional urban areas of voivodship capitals	**20**	0	10	0	2	0	0	0	0	1	0	13	0.39	33
2	external zones of fua's of voivodship capitals	0	**85**	12	78	28	44	10	2	6	0	0	180	0.68	265
3	functional urban areas of subregional centres	4	0	**44**	0	7	0	0	0	0	0	0	11	0.20	55
4	external zones of fua's of subregional centres	0	8	3	**75**	9	53	26	6	21	0	0	126	0.63	201
5	multifunctional urban centres (other)	0	0	5	8	**127**	0	0	1	1	0	0	15	0.11	142
6	communes with developed transport functions	0	0	0	14	18	**34**	16	32	23	0	0	103	0.75	137
7	communes with other developed non-farming functions (tourism and large-scale functions, including mining)	0	2	0	18	17	13	**102**	30	39	0	1	120	0.54	222
8	communes with intensive farming functions	0	0	0	5	3	62	0	**388**	38	0	0	108	0.22	496
9	communes with moderate farming functions	0	1	0	35	21	118	33	144	**313**	0	0	352	0.53	665
10	extensively developed communes (with forests or nature protection areas)	0	0	0	7	9	15	112	35	84	0	0	262	1.00	262

The problem at hand (8)

Typology of some 2 500 Polish municipalities for definite planning purposes

The **weights of variables** in [one of] the „best” obtained partitions (summing to 1):

1. Population numer: 0.382

2. Overbuilt area: 0.329

3. Share of transport related areas: *0.022*

4. Population density: 0.000

5. Share of agricultural land: 0.019

6. Share of overbuilt areas: 0.002

7. Share of forest areas: 0.004

8. Share of population over 60 years of age: 0.001

9. Share of population below 20 years of age: 0.003

10. Birthrate for the last 3 years: 0.001

11. Migration balance for the last 3 years: *0.040*

12. Average farm acreage indicator: 0.011

13. Registered employment indicator: *0.044*

14. Registered businesses per 1 000 inhabitants: *0.057*

15. Employment-based average business magnitude indicator: 0.006

16. Share of businesses from manufacturing and construction: 0.012

17. Number of pupils per 1 000 inhabitants: 0.001

18. Number of students of over-primary schools per 1 000 inhabitants: *0.034*

19. Own revenues of municipality per inhabitant: 0.010

20. Share of revenues from personal income tax in own communal revenues: 0.023

21. Share of social care expenses in total communal budget: 0.000

The problem at hand (9: Conclusions)

Typology of some 2 500 Polish municipalities for definite planning purposes

-- acceptable qualitative re-establishment of the original partition;

-- quite distinct reference to („correlation with”) the urban-rural axis;

-- effective identification of some of the „special types” (also beyond the original partition), but not all of them;

-- implication that the unidentified special types might be „artificial”, requiring yet other variables, or even „nonexistent” (see the limit of 10 types);

-- important additional information (e.g. variable weight);

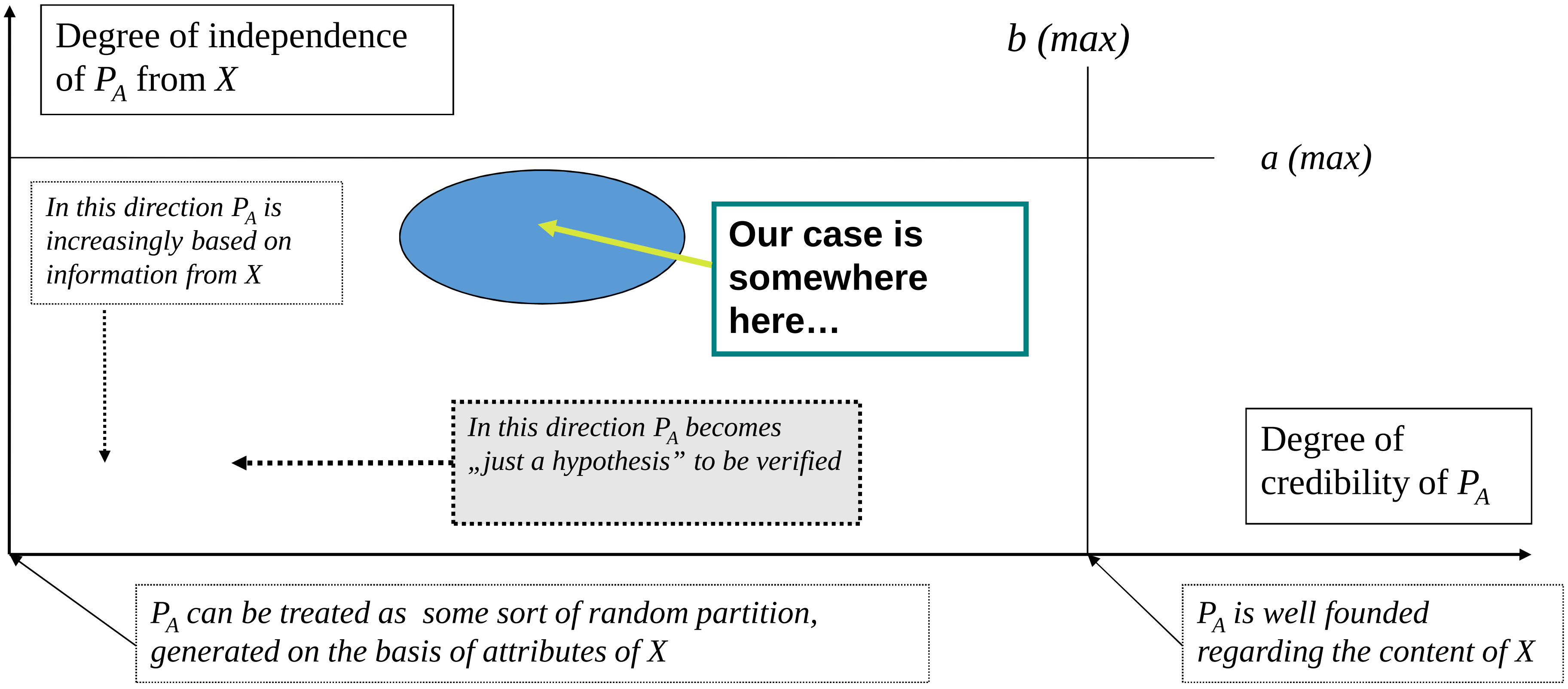
-- implication that a better categorisation might be obtaned.

Another (apparently similar) example…

Exemplary results for P_{A1}: best k-means based partition

Initial administrative categories for one province in Poland (Masovia, the capital province):	***Calculated categories:***		
	1	*2*	*3*
1. Urban municipalities	**30**	**0**	5
2. Rural municipalities	**0**	**224**	4
3. Urban-rural municipalities	2	35	**14**

The understanding, again… (why…?)



Degree of independence of P_A from X

b (max)

a (max)

In this direction P_A is increasingly based on information from X

Our case is somewhere here…

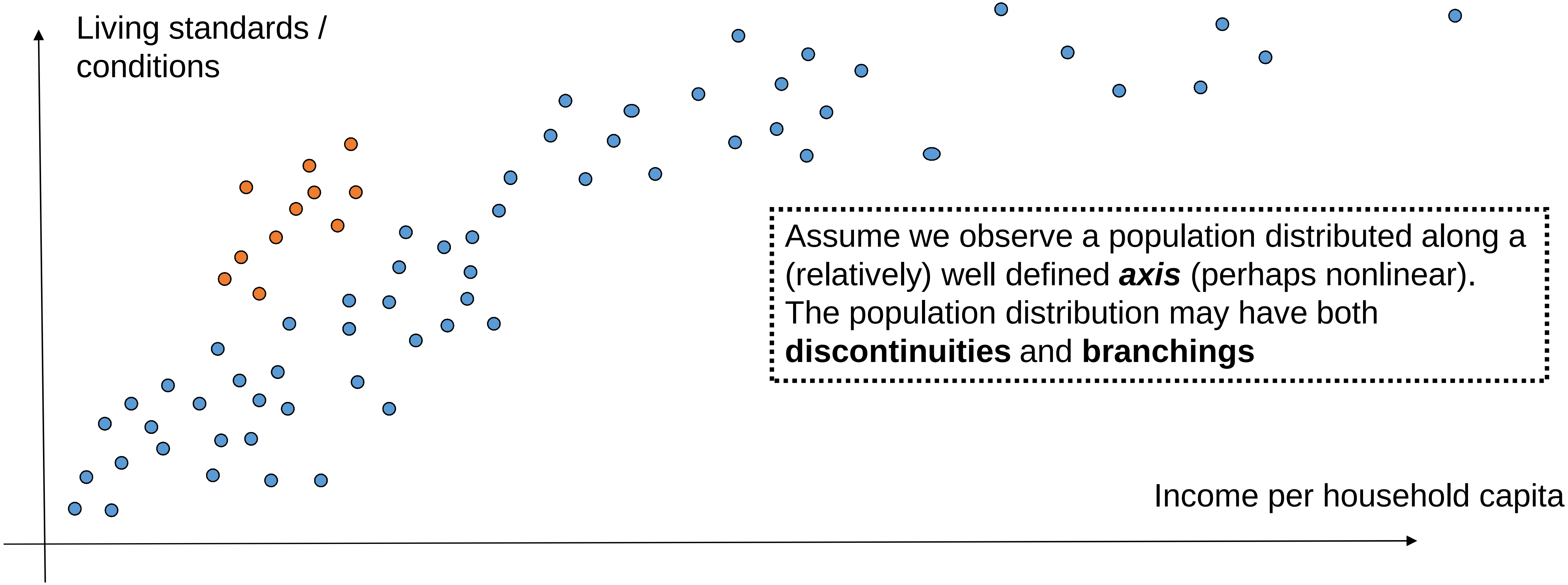
In this direction P_A becomes „just a hypothesis” to be verified

Degree of credibility of P_A

P_A can be treated as some sort of random partition, generated on the basis of attributes of X

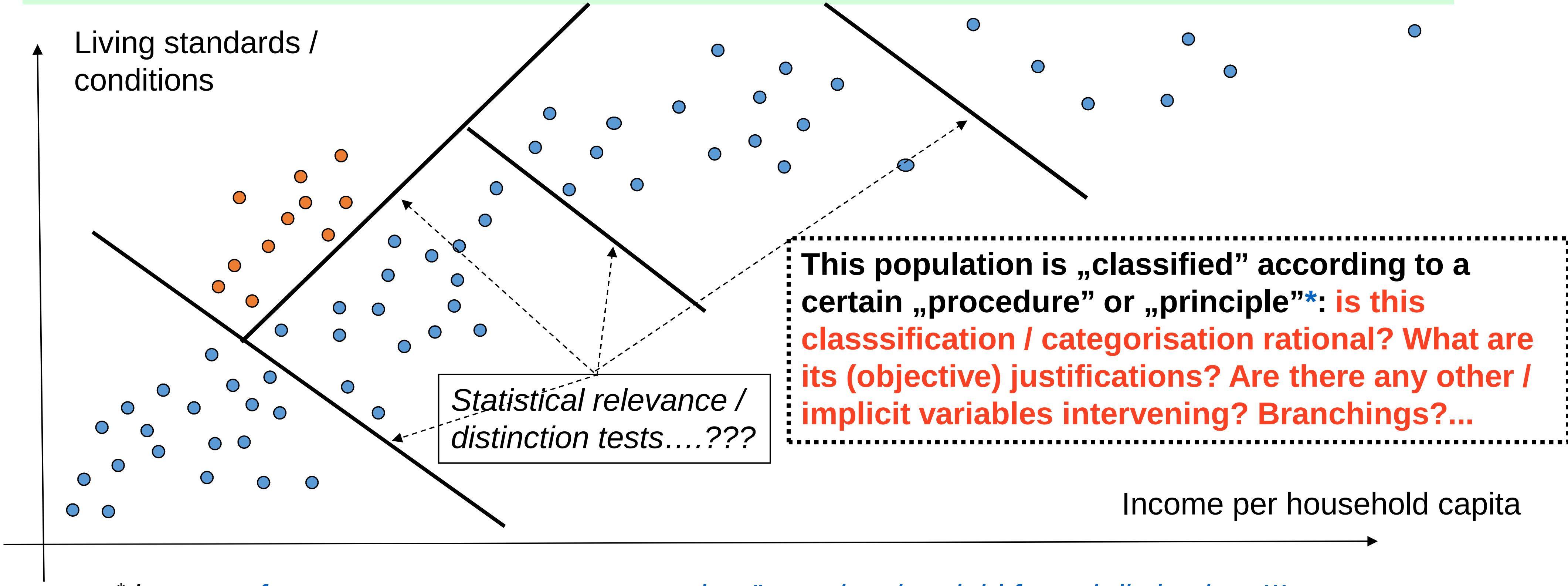
P_A is well founded regarding the content of X

A broader picture (1)



Assume we observe a population distributed along a (relatively) well defined ***axis*** (perhaps nonlinear). The population distribution may have both **discontinuities** and **branchings**

A broader picture (2)



Statistical relevance / distinction tests….???

This population is „classified” according to a certain „procedure” or „principle”*: is this classsification / categorisation rational? What are its (objective) justifications? Are there any other / implicit variables intervening? Branchings?...

* *in case of poverty assessments a „procedure” may involve rigid formal distinctions!!!*

Some final conclusions

- The quantitative results obtained are, in general, quite promising:

a. the partitions obtained for the „well-justified" cases are very close to the original ones,

b. the differences are almost always telling in terms of both interpretation and methodology, **in some cases showing „better" characteristics than the original partition**

c. when solutions obtained are (perceptibly) far from the original partition, hints can be formulated on the missing information (variables) and either the ways to complement it, or the inconsistency thereof

c. the parameters obtained can be effectively used for other similar data sets.

The approach proved to be numerically feasible for small cases but computationally cumbersome for larger ones (hence: further work, especially on parallelisation, but also on *better search procedures* – enhancement of search effectiveness)

The work continues on both technical and substantive sides, e.g. in the direction of ***outlier detection***, which turned out to be a specially promising ground

Thank you very much for *listening*
(that is – if you have *listened*…),
and also
for patience (if you really did this, I mean: *listened*)…