Half-Taxi Metric

in Compositional Data Geometry *rcomp*

Katarina Košmelj and Vesna Žabkar

Biotehnical Faculty, University of Ljubljana, Slovenia;
katarina.kosmelj@bf.uni-lj.si

Faculty of Economics, University of Ljubljana, Slovenia;
vesna.zabkar@ef.uni-lj.si

I. INTRODUCTION

Advertising expenditure (ADSPEND) includes the following advertising media

- **Electronic** (Radio, TV)
- **Print** (Press, Outdoor)
- **Online** (recently, supported by Internet)

Data for 17 countries for 1994-2008 (Source: Euromonitor, 2009)
***stable* countries** (ADSPEND/GDP approx constant (0.7%); most developed European Union countries and two Baltic countries

The data for ADSPEND are presented in the local currency and is not comparable between countries. Therefore it can not be analyzed in the original form; **a transformation needed.**

Proportions for each country in each year

Austria (%)	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Electronic	37.6	35.1	33.7	35.0	34.2	33.6	33.8	33.3	32.2	32.4	33.2	32.1	31.9	31.6	31.4
Print	62.4	64.9	66.3	65.0	65.8	66.4	66.2	66.2	66.6	67.1	65.8	66.6	66.5	66.5	66.4
Online								0.5	1.2	0.5	1.1	1.3	1.7	1.9	2.2

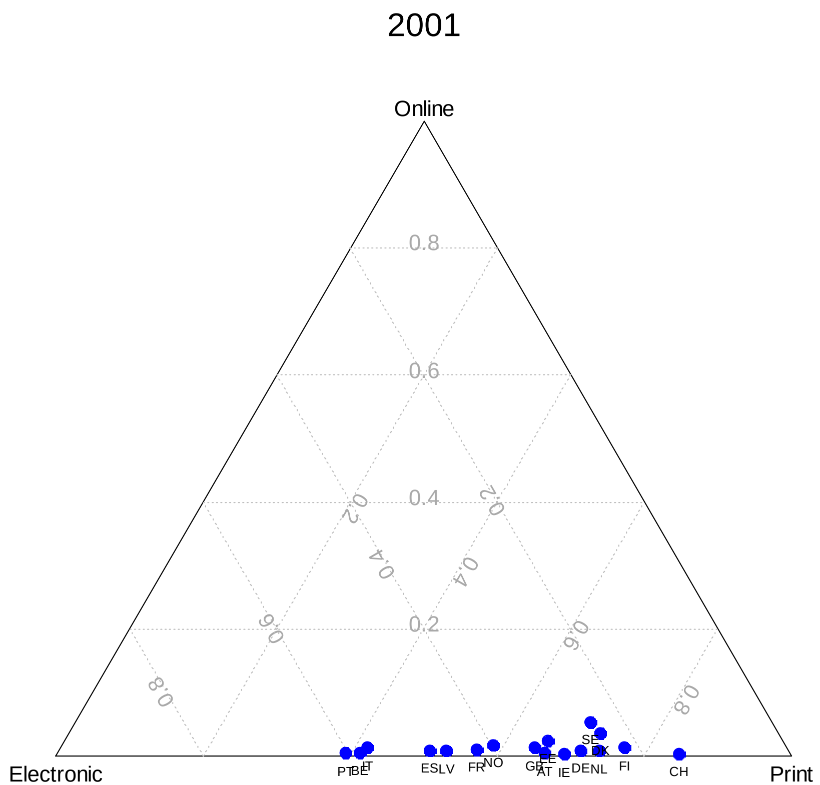
Online component

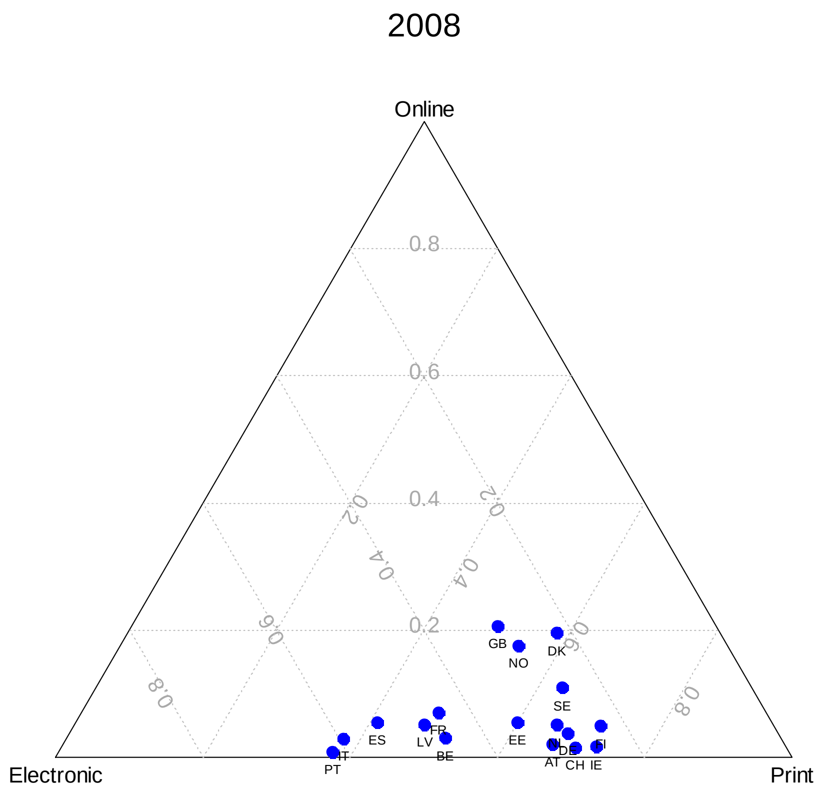
Online	Country	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Austria	AT								0.5	1.2	0.5	1.1	1.3	1.7	1.9	2.2
Belgium	BE					0.1	0.4	0.7	0.6	0.6	0.8	1.4	1.8	2.5	2.8	3.1
Switzerland	CH					0.2	0.3	0.6	0.5	0.5	0.8	0.9	1.1	1.4	1.6	1.7
Germany	DE					0.1	0.4	0.8	1	1.4	1.6	1.7	2	2.9	3.5	3.9
Denmark	DK								3.8	5.4	5.9	6.5	7.6	15.3	18.1	**19.6**
Estonia	EE					0.4	0.6	1.9	2.5	2.5	3.1	2.9	3.5	4.9	5.5	5.6
Spain	ES					0.1	0.3	0.9	1	1.3	1.4	1.6	2.5	4.3	5.1	5.6
Finland	FI			0.1	0.2	0.3	0.6	1	1.4	1.4	1.6	2	3	3.8	4.4	5
France	FR				0.1	0.2	0.9	1.5	1.1	1	1.3	1.6	3.4	4.6	6.3	7.2
Un. Kingdom	GB				0.1	0.2	0.5	1.3	1.4	1.6	2.9	6.2	10	14.5	17.7	**20.7**
Ireland	IE							0.3	0.3	0.4	0.5	0.7	1.1	1.5	1.7	1.8
Italy	IT					0.1	0.4	1.7	1.4	1.3	1.3	1.3	1.6	2.3	2.8	3.1
Latvia	LV							0.3	0.9	1.2	1.9	1.8	2.5	4.4	5	5.3
Netherlands	NL						0.6	1	0.9	0.9	1.2	1.9	2.8	3.8	4.5	5.2
Norway	NO							2.3	1.8	1.9	2.1	2.6	10.2	13.6	16.1	**17.7**
Portugal	PT					0.6	0.5	0.5	0.6	0.6	0.5	0.4	0.5	0.8	0.9	1
Sweden	SE				0.4	1.3	3.1	5.6	5.5	7.2	8	10.9	14.6	11.4	11.1	11

1994-1995: Online did not exist yet

1996 onwards: Online develops in time; near zero values and no data
Some values are not collected/reported; see DK before 2000, NO before 2000.

2001: the first year with Online data for all countries.


2001
Online
Electronic
Print
0.8
0.6
0.4
0.2
PT
BE
IT
ES
LV
FR
NO
GB
EE
AT
IE
DE
NL
SE
DK
FI
CH


2008
Online
Electronic
Print
0.8
0.6
0.4
0.2
GB
NO
DK
SE
ES
FR
LV
BE
EE
IT
PT
AT
CH
IE
DE
FI

OBJECTIVES

Identify structural changes in the components.

For which countries is an increase in Online made on the account of Print, on the account of Electronic or on the account of both?

II. STATISTICAL ANALYSIS

Compositional data: the spurious correlations are induced by the constant sum constraint.

R package: compositions

***acomp* (Aitchison composition)**
Distance is based on the **relative scale**: 1 and 2 are as far as 10 to 20)

***rcomp* (Real composition)**
Distance is based on the **absolute scale difference**: 1 and 2 are as far as 51 and 52
Difference is 1 percentage point (1 pp)

Which geometry is suitable for our problem?

- *acomp* geometry overemphasizes components with near zero values for Online;
- absolute scale of interest

K.G. van den Boogaart, Applied Statistics, 2009

We can analyse a dataset of portions with classical multivariate methods if ALL of the following assumptions are TRUE

a) data normalized to 1

b) there is only one type of measurement units reasonable

c) all possible/thinkable components are in the dataset

d) absolute difference on percentage is meaningful

rcomp geometry is acceptable for our problem

Notation: $n \geq 2$

$$\mathbf{x} = [x_1, x_2, ..., x_n] \qquad x_i \geq 0 \qquad \sum_i x_i = 1$$

$$\mathbf{y} = [y_1, y_2, ..., y_n] \qquad y_i \geq 0 \qquad \sum_i y_i = 1$$

The set of compositions is a $(n-1)$-dimensional simplex **with the boundary**.

Which distance is suitable for the *rcomp* geometry?

Approach 1: similarity coefficient

MILLER, W. E. (2002): Revisiting the geometry of a ternary diagram with the half-taxi metric. Mathematical Geology, 34(3), 275-290.

Miller defines a similarity coefficient

$$s(\mathbf{x},\mathbf{y}) := \min\{x_1, y_1\} + \min\{x_2, y_2\} + \ldots + \min\{x_n, y_n\}$$

Taking into account the expression

$$\min\{a,b\} = \tfrac{1}{2}\left(a + b - |a - b|\right)$$

and the fact that compositions are closed to 1, it follows

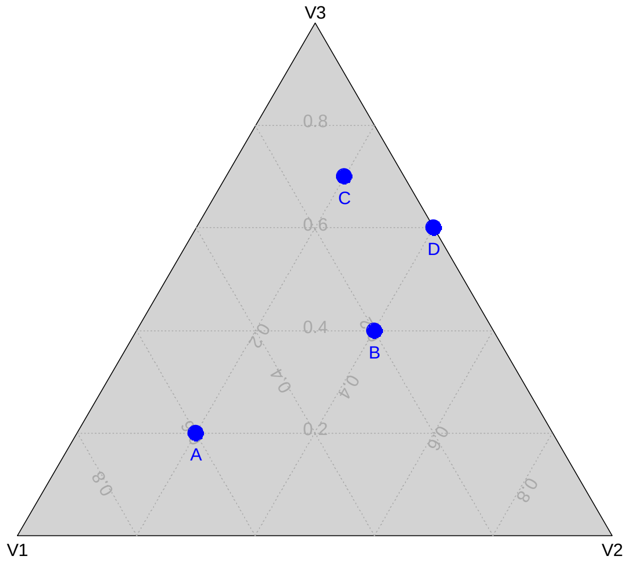
$$s(\mathbf{x},\mathbf{y}) = 1 - \frac{1}{2}\left(|x_1 - y_1| + |x_2 - y_2| + \ldots + |x_n - y_n|\right)$$

The complimentary form is a dissimilarity coefficient:

$$d(\mathbf{x},\mathbf{y}) := 1 - s(\mathbf{x},\mathbf{y}) = \frac{1}{2}\left(|x_1 - y_1| + |x_2 - y_2| + \ldots + |x_n - y_n|\right)$$

.

- **Half of the standard taxi ("Manhattan") distance**
- **Geometric interpretation**: it presents the shortest path between points **x** and **y** on the triangular coordinate system



Manhattan distance	A	B	C
B	0.8		
C	1.0	0.6	
D	1.2	0.4	0.4

Approach 2: heuristic approaches

HAJDU, L. J. (1981): Graphical Comparison of Resemblance Measures in Phytosociology. Vegetatio, v. 48, 47-59.

- SIM7 (Hajdu)
- percentage similarity of distribution
- relativized Czekanowski coefficient
- relative absolute value function
- Renkonen, 1938; Whittaker, 1952, Orloci, 1973

Approach 3: based on the theory of normed metric spaces

Let us choose a norm $\|\cdot\|$ on R^n which is "suitable" for the problem under study. This norm induces a **norm metric** $n(\mathbf{x},\mathbf{y}) := \|\mathbf{x}-\mathbf{y}\|$ on R^n.

Let M be a subset of R^n, with the property that any two points are connected by a path of finite length. (The finiteness of a path length does not depend on the choice of the norm).

In the subset M we define the **intrinsic metric** (also called **length metric**) $d(\mathbf{x},\mathbf{y})$ as follows:

$$d(\mathbf{x},\mathbf{y}) := \inf\{L(\mathbf{a}) \mid \mathbf{a}(t) \text{ is a path within } M \text{ from } \mathbf{x} \text{ to } \mathbf{y}\}$$

$L(\mathbf{a})$ is the path length defined by the norm metric $n(\mathbf{x},\mathbf{y})$.

The intrinsic metric is defined as the infimum of lengths of all paths from one point to the other within M.

FACT: If M is a convex set, then its length metric agrees with the original norm metric: $d(\mathbf{x},\mathbf{y}) = n(\mathbf{x},\mathbf{y})$.

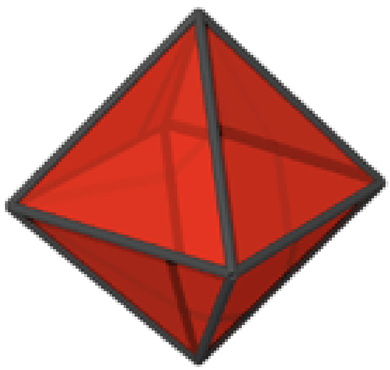
Application to compositional data

The **unit sphere** $S = \{x \in R^n \mid \|x\|_1 = 1\}$ in l_1-normed space is the **surface of a cross-polytope**.

The compositional data sample space $M = \left\{ x \mid x_i \geq 0, \sum_{i=1}^{n} x_i = 1 \right\}$ is a **simplex** and **is a part (a face) of this cross-polytope.** This simplex is a **convex** set in R^n.

Illustration for $n = 3$:

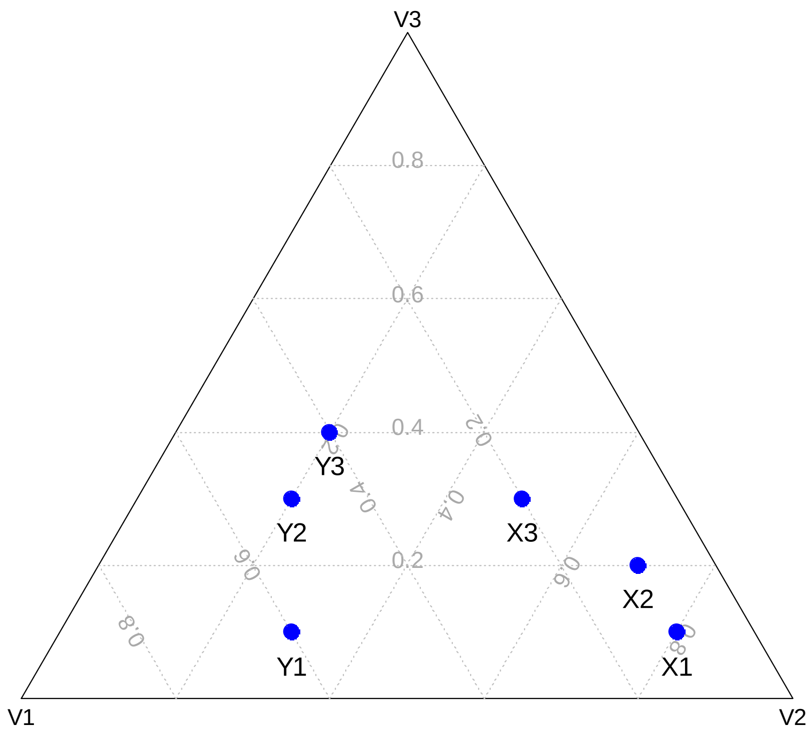
- the unit sphere l_1-normed space is the surface of an octahedron
- the compositional data sample space is one of its triangles



Therefore, for analysis of compositional data in *rcomp* geometry

- **the l_1-norm can be considered as the most natural choice of a norm,**
- **and hence its norm metric (taxi distance) as the most natural choice of a metric**

DISTANCE BETWEEN TWO TIME TRAJECTORIES



d_t=distance at a time point t,

w_t=weights at time t

Distance between two time trajectories

$$D(\mathbf{X},\mathbf{Y}) := \sum_{t=1}^{T} w_t \cdot d_t$$

d_t …. Manhattan distance

w_t …. internet users per ‘000

III. RESULTS

We analyzed the data from 2000 onward

Online	Country	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Austria	AT								0.5	1.2	0.5	1.1	1.3	1.7	1.9	2.2
Denmark	DK								**3.8**	5.4	5.9	6.5	7.6	15.3	18.1	**19.6**

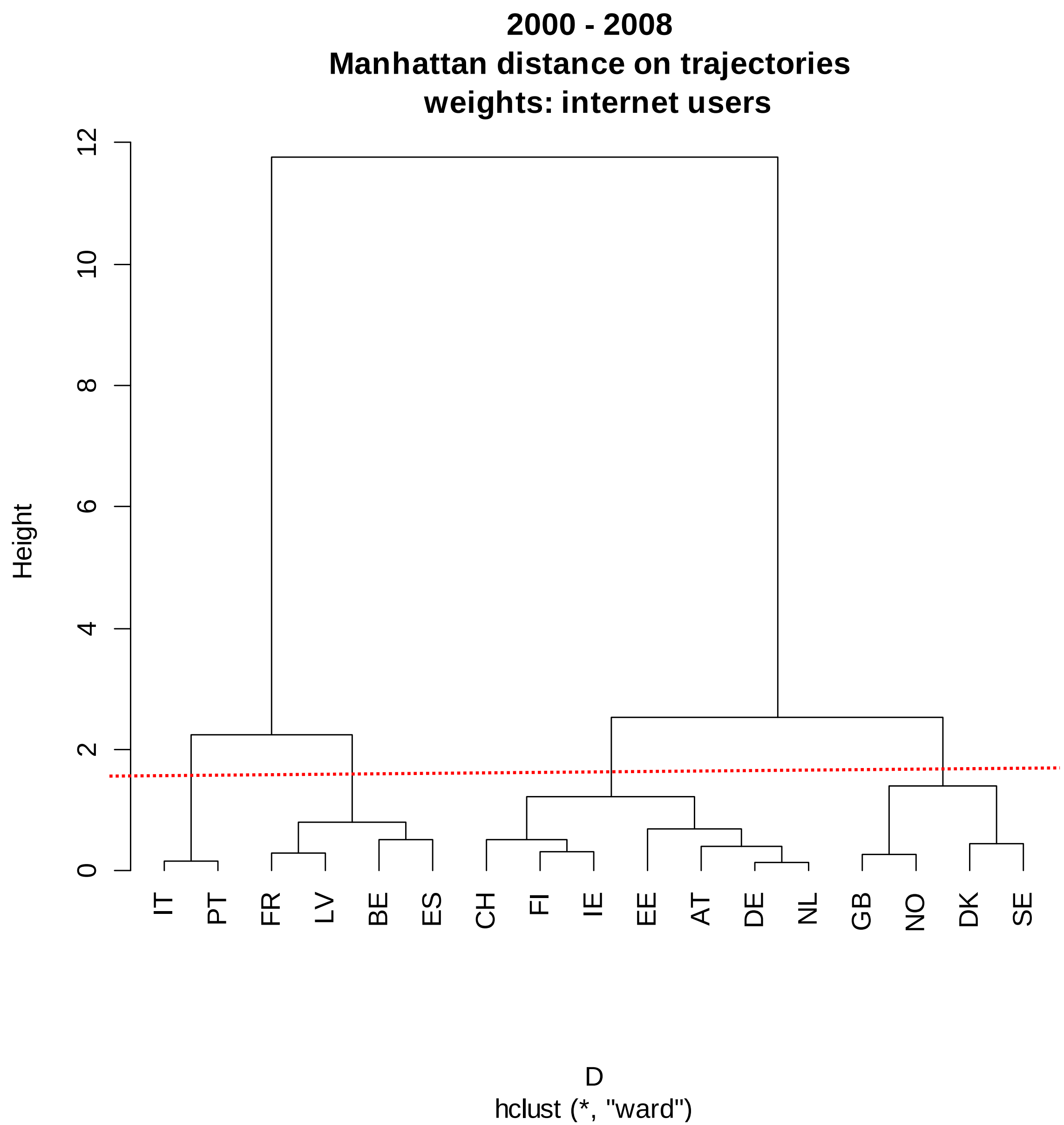
Two values imputed:

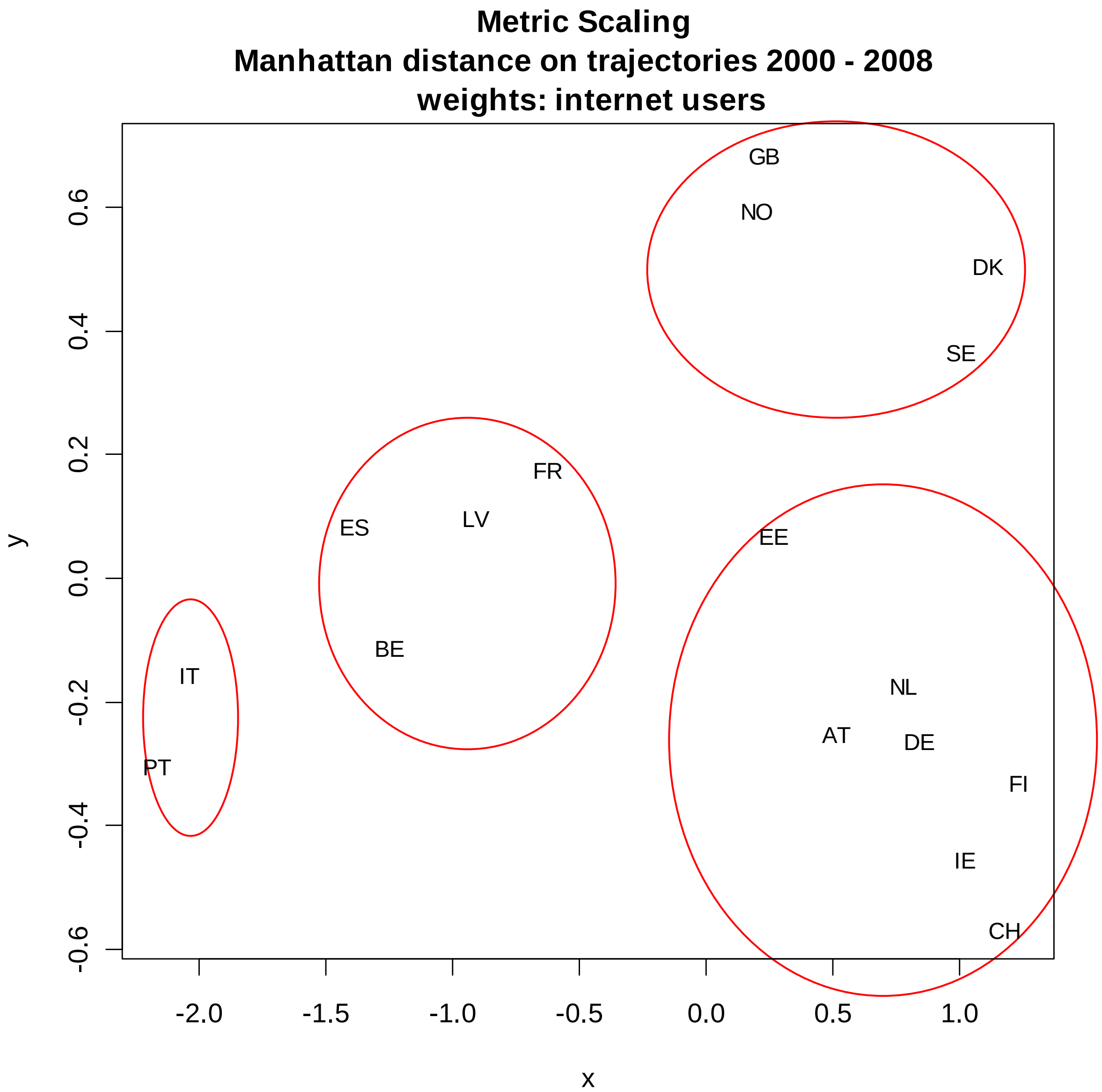
AT: 0

DK: ???

w_t …. internet users per '000

	2000	2001	2002	2003	2004	2005	2006	2007	2008
w	0.253	0.305	0.422	0.485	0.532	0.564	0.602	0.635	0.664


2000 - 2008
Manhattan distance on trajectories
weights: internet users
Height
0
2
4
6
8
10
12
IT
PT
FR
LV
BE
ES
CH
FI
IE
EE
AT
DE
NL
GB
NO
DK
SE
D
hclust (*, "ward")



x axis:
Left: High context cultures
Right: Low context cultures

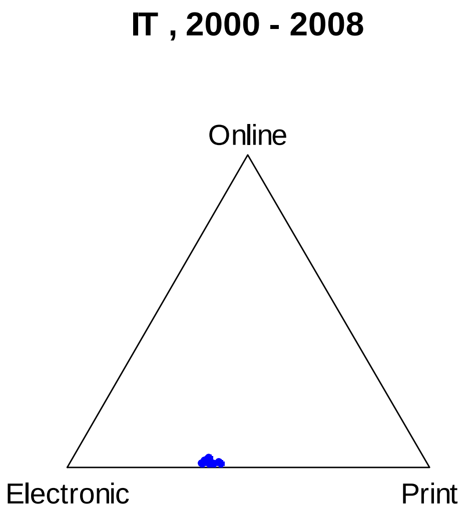
y axis:
Bottom: no change in time
Top: change in time O↑

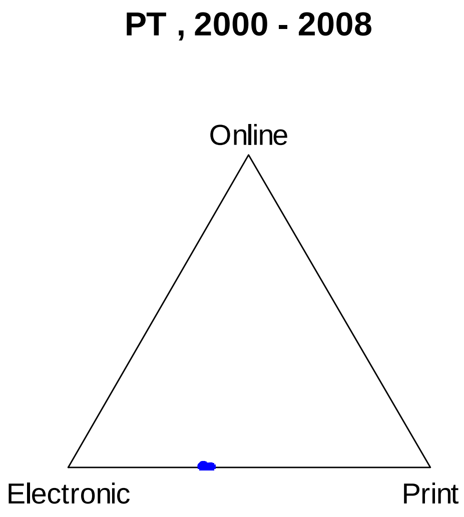
High-context cultures have closer and more familiar contacts with each other; their preferred mode of communication is more informal, indirect, and often based merely on symbols or pictures."

"In low-context cultures, individuals have less personal contact with each other; the communication must be very detailed, formal, very explicit, communicated in a direct way, often by way of written texts."

Cluster 1: IT, PT

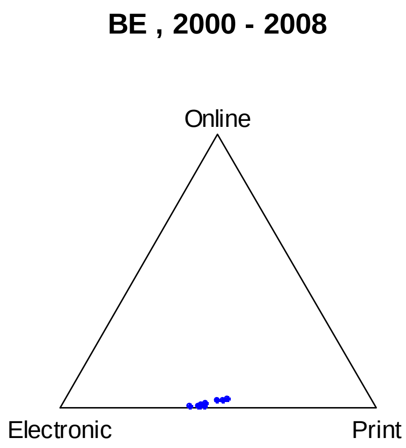
Stationary, Electronic Dominant (E≈0.6 , P≈0.4)

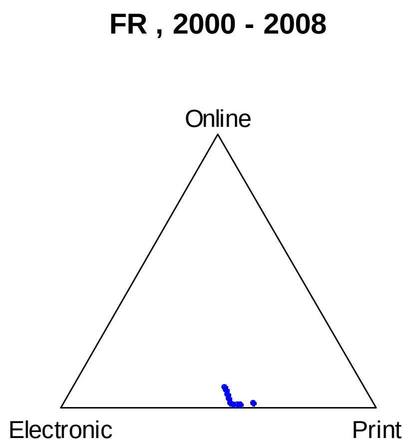




Cluster 2: FR, BE, ES, LV

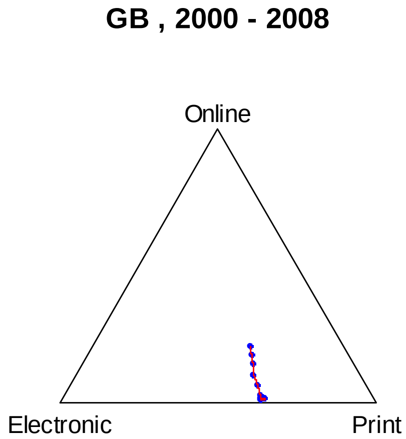
Electronic and Print approx 1/2 , modest increase in Online

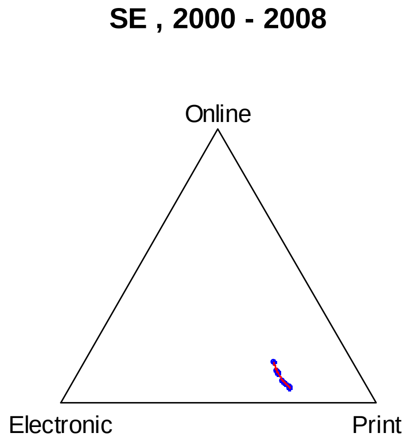




Cluster 3: GB, NO, DK, SE

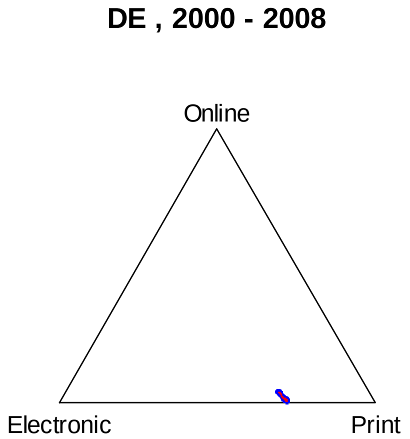
Significant increase in O (up to 0.2) on the account of P

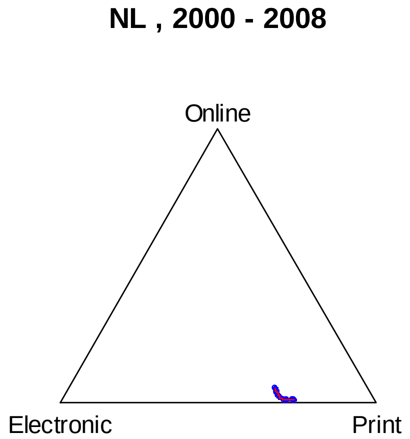




Cluster 4: CH, IE, FI, DE, NL, AT, EE

Print Dominant, modest increase in E and O on the account of P





IV. **CONCLUSIONS**

- **It is well known that problems can arise when treating compositional data with conventional statistical techniques. It is not possible to distinguish between spurious effects caused by the constant sum constraint and the effect attributable to the process under study;**

- ***acomp*** **geometry is to be used**

- ***rcomp*** **geometry is rarely applicable in practice. Severe conditions are to be satisfied for its use.**

    - **Zero and near zero values do not cause any problems**

    - **Manhattan distance is the most natural since the compositional data sample space is a part of the unit sphere in l_1-normed space.**

        - **Results on advertising expenditure detect structural changes in time, in view of the newer Online component.**