SELF ORGANIZING MAPS: A CLUSTERING NEURAL METHOD FOR URBAN ANALYSIS

Lorena Franzini*, *lorena.franzini@polimi.it*
Paola Bolchi* *paola.bolchi@polimi.it*
Lidia Diappi* *lidia.diappi@polimi.it*
**Dept. Architecture and Planning, Polytechnic of Milan, Milan, Italy*

RÉSUMÉ. La recherche concerne l'expérimentation d'une méthode de traitement des données basée sur les Réseaux Neuraux et donc sur une approche cognitive de nature connexionistique. Le propos c'est d'explorer la structure de relation entre indicateurs urbains et connaître les liens structuraux des systèmes urbains.L'étude a analysé les connexions présentes dans un ensemble de variables pour la ville de Milan, divisée en 144 zones statistiques.La base de données pour Milan a été analysée à partir d'une définition de soutenabilité urbaine, interprétée comme interaction positive entre les systèmes social, économique et environnemental, et de déterminer une série de indicateurs pour exprimer les relations entre les trois systèmes.Précédemment la même base de données a déjà été analysée avec une méthodologie de mesure de risque basée sur une approche de seuil et avec Reseaux Neuraux Auto-Associés. Dans ce papier on présente une nouvelle des Self Organizing Maps (SOM).Les SOM mettent en oeuvre un processus d'organisation spatial selon lequel les unités de input sont classifiées selon leur caractéristiques urbaines. Le réseau partage logiquement l'éspace de input en " cluster ", où prévalent les composantes principales qui différencient les données en entrée.Les SOM sont reseaux qui évoluent avec un apprentissage non-supervisioné et auto-organisatif, en faisant fonctionner les unités de input en compétition selon un principe de " winner unit ".L'application de cette typologie de reseaux permet de construire des cartes de Milan en maintenant la désagrégation spatial en zones statistiques.Les résultats des SOM ont permis de réaliser plans de " clustering " où les zones sont identifiées soit par leur forts similarités soit par la position qu'elles occupent dans la matrice de output (de laquelle la dimension doit être définie au début du processus neural).

ABSTRACT. The research focuses on the testing of a cognitive method for data processing based on Neural Networks. The approach followed is connectionistic and it is applied with the intention to investigate the relationships among urban indicators and understand the structural links of urban systems.The research analyzes the connections in a set of variables for the city of Milan subdivided into 144 statistical areas.The database of Milan is investigated starting from a specific definition of urban sustainability, meant as positive interaction between social, economic and environmental systems, and then collecting a set of indicators aimed to express the relationships between the three systems.During previous researches the same database has been investigated with a methodology of risk measurement based on a threshold approach and with the application of Self-Reflexive Neural Networks. This paper presents a new kind of application in which Self Organizing Maps (SOM) are experimented. SOM carry out a process of spatial organization in which the spatial units are classified on the basis of their own urban characteristics. Network splits up the input records into clusters, in which the main components that differ input data become prevalent.SOM are a typology of Networks that evolves through an unsupervised and self-organized learning process, by elaborating in competition the input units and referring to a "winner unit" concept.

The application of this kind of Networks makes possible to construct maps of Milan by maintaining the spatial disaggregation in statistical areas.The outcomes of SOM also allow to obtain clustering maps where areas are identified both by their strong similarity and by their position in the output matrix (whose dimension has to be determined at the beginning of the neural process).

***Mots- clés**: Réseaux Neuraux, soutenabilité urbaine, auto-organisation, clustering, classification flou.*

***Key words**: Neural Networks, urban sustainability, self-organization, clustering, fuzzy classification.*

Introduction

The research is aimed to test a classification method based on *Self Organizing Maps* (SOM) applied to the urban system.

The study has been carried out on a Data Base of 56 social, economical and environmental indicators referred to the city of Milan, spatially partitioned into 144 zones.

Indicators and amount of zones are such to allow many potential investigations on the complex structure of the Milanese urban system. Artificial Neural Network (ANN) possibilities are still largely unexplored, specially when applied to urban studies. For this reason they have been chosen as field of interest by the authors.

The importance of ANN lies not only in being a new sophisticated methodology and technology for data handling. Above all it is a new way to see phenomenon, arisen from the emerging similarity between the evolution of the knowledge in various scientific fields and cerebral learning processes, studied in depth and formalized by the “neuroscience”.

1. Information and signals mapping

SOM have been developed mainly by Kohonen between 1979 and 1982 (Breda, 1999).

Kohonen researches (1995) have been motivated by the possibility to represent knowledge of specific categories as geometrically organized *feature maps*.

The main problem in handling large amount of information is to find structures for memorizing, classifying and representing data as efficient as the human brain.

The human thinking when processing perceptive and subconscious information, tends to squeeze them forming *reduced representations* of the most relevant facts, without loosing awareness of their interrelationships.

In the brain, areas are specialized and connected to the different sensorial modalities (fig.1). The signal response models follow a topological order. For instance, in the *tonotopic map* of the auditory cortex the spatial sequence of the cellular response is the same as the heard frequencies.

Brain ability to create cellular nets receiving signals depends on a geometrical order with different abstraction level. Brain maps are genetic, nevertheless several empirical evidences show that also the experience can modify and enrich the map forming. A classical example occurs when a sensorial organ is damaged and the corresponding cells are immediately dedicated to other uses.

The same principle can also introduce hierarchical orders by assigning different sub-areas to different abstract information levels.

The main task matched by SOM is to run a self-organizing process that, through mathematical algorithms, creates maps similar to the brain ones.

SOM create two-dimensional ordered maps from multidimensional signals, maps dimension corresponds to the characteristics of main input signals.

The most relevant examples are about acoustic frequencies, language phoneme, colors (hue and saturation), textures and organization of geographical locations (Kohonen, 1995). In other very important maps category, localization describes some contextual characteristic existing in input symbols strings. For instance, if strings represent words of some natural language, such maps can identify the words as nouns, verbs, adverbs or adjectives.

Why do two-dimensional output maps result the more efficient ones?

From a merely physiological point of view it can be presumed that, when differences between nearby groups are little, communication between similar cells will be easier. Moreover, as biological components are not as stable and well defined as the technical ones, the partition reduces the risk of unwanted interference and communications between brain’s parts.

In theoretical models, the neural function seems to form clusters and hierarchies, and it is well-known that clustering is the first step towards abstraction. The representation of these hierarchically ordered concepts seems to arise automatically from these processes.

It is surprising how, in neural models, so little attention has been paid so far to spatial order, that, instead, is a key element to understand the systems internal organization.

2. Self-Organizing Maps

Self Organizing Maps are two-dimensional neural networks whose nodes arrange multidimensional input signals. These networks apply a competitive and unsupervised learning process, meaning that no output (or target) model is given, but it is the network itself that puts signals in order, according to a two-dimensional meaningful coordinate system.

SOM are also defined auto-poietic as target is dynamic and keep changing during the whole learning.

SOM architecture comprises two layers:

- the one-dimensional input layer $\boldsymbol{X}=(x_1, x_2, \ldots, x_N)$: a record formed by the input variables and fully connected to the Kohonen matrix (each unit in the matrix has in input each unit of the input layer);
- the two-dimensional output layer, known as Kohonen matrix: its nodes are placed to form a matrix $M=M_R \cdot M_C$ with M_R row and M_C column.

This matrix is formed by units, regularly organized in the space and evolves during the training following a spatial organization process of the data, named Feature Mapping.

At each unit in the matrix it is associated a weights vector $W_r=(W_{r1}, W_{r2},\ldots, W_{rN})$ (*codebook*) corresponding to the connections to the input units.

Before the training weights vectors are initialized with small random values, assigning different vectors to each different matrix unit.

Learning runs through many cycles and during the process input vectors X are showed to the network randomly or sequentially.

For every vector X it is found the matrix unit whose *codebook* has the shortest distance D_s from X: this unit is nominated *winner unit* (WU). The distance can be calculated in several ways, the simplest is the Euclidean D_r.

$$D_r = \sqrt{\sum_{i=1}^{N} (x_i - w_{ri})^2}$$

This updating process takes then place not only on the WU, but also on its nearby units. In this way, when the learning phase is through, neighboring units will have “similar” *codebooks* and the relative position of the units in the map represents an ordered concept.

The weights in the neighborhood are not updated by the same amount of the WU, but depending on a function having a maximum in coincidence with the WU and decreasing moving away from it (linear, “Mexican hat”, Gaussian or sine). The function width defines the neighborhood dimension, and it is decreased in each cycle till to reduce the neighborhood to the single WU (fig. 2).

Other parameters to be carefully calibrated are:

- the minimum and maximum function width;
- the neighborhood shape, square or rhomboidal;
- the weights correction factor during the cycles.

It can be observed that the random weight initializing factor can have a light influence on the results.

When the net is trained more times using same data set and parameters, two important things can be observed: first, the layout can undergone rigid transformations (rotations, mirroring, …) as it is not defined any initial absolute reference system; second, similarities between *codebooks* imply that little characterized vectors can be mapped each time to different unit.

At the end of the training phase, the net maps each input vector to the output unit that has the minimum Euclidean *codebook* distance.

The SOM attitude to “classify” makes possible to perform a *mapping* according to two main goals (Breda, 1999):

- Clustering: the net performs a logical division of the input space into regions (cluster), associating a point in the N-dimensional input space to the two-dimensional output matrix. In the dimension reduction process the principal components discriminating data are dominant.

- Self-organization: before the training the weights vectors topology depends only on the initializing criterion: if random the weights will be casually organized into their hyper-cube. The learning criterion tends to move the weights vectors toward input vectors seen during the training. This behavior affects not only the winner unit vector, but also its neighborhood according to a decreasing function.

3. SOM application in urban analysis

SOM experimentation has been carried out using a square output matrix. Results have been interpreted and analyzed in three different ways:

- changing the dimension of the output matrix (units number);
- reading clusters profiles through their *codebook*;
- creating a color hatched plot of the zones, based on their belonging to the output matrix.

During the first experiments the units number in the output matrix has been changed from 9 (3x3) to 49 (7x7).

Changing the matrix dimension, the net makes a new profiles classification according to a dynamical interpretation.

Obviously the number of zones in each cluster decreases as the matrix dimension increases, rising a meaningful quantitative differentiation. Some zones will be alone in their cluster, but the one with stronger similarity remain grouped together.

Comparing maps related to different matrix dimension some interesting constant can be observed:

- city center is always in a cluster of his own, placed in a corner of the matrix;
- the more characterized zones are placed at the extremity of the matrix also when changing the matrix dimension.

Results have been read analyzing the 25 unit matrix experiment.

The existence of a random factor during the weight matrix initialization causes some little differences in the results, when the training is repeated on the same data set with the same parameters, some zones can be mapped in different clusters.

This kind of behavior “bias” to some extents the results and their possibility to be compared. In the same time, a general repetitiveness in the clustering through the experiments enforces their meaning and reliability.

As second step the output layer has been analyzed considering the *codebook* vectors, then describing the prototypical profiles of each cluster.

Representing all the *codebooks* in a graph (fig. 4) the *cluster* strongly characterized by some variables can be individuated. These are the most meaningful to point out some urban features.

In the same way it can be charted the minimum and maximum values of each variable in the *codebook* set, so representing how the net perceives them.

The zones, considered as hyper-points, having larger distance from their *codebook* (fig. 5), are the ones “forced” by the net into their cluster. For at least one variable they are far from the output description given by the algorithm.

As further and final step color maps have been realized to visualize the spatial layout of the resulting cluster. This enables to locate into the urban area zones and their classification, so to verify if zones in the same cluster are also spatially clustered or if similar urban characters are scattered in town.

The following elements need to be defined to run a SOM application.

1. Net architecture

Input Unit	Number of units in the input vector (56)
K Units	Number of units in the output matrix (from 9 to 49, depending on the simulation)
K Rows	Number of rows in the output matrix (from 3 to 7, depending on the simulation)
K Cols	Number of columns in the output matrix (from 3 to 7, depending on the simulation)
K Dimension	Output matrix dimensions (2)
K Topology	Output matrix space topology (Euclidean)
N Topology	Winner unit neighborhood space topology (square)

2. Parameters

N function	Parameter defining the function to update the units connections in the WU neighborhood (Gaussian)
Alpha Max	Maximum width for the *N function* (1)
Alpha Min	Minimum width for the *N function* (0)
Alpha Inc	Factor reducing *Alpha Max* in each epoch (0.01)
Set Weight	Maximum weight value during the initialization
Alpha W Func	Input/output weights correction function (constant)
Alpha W Max	Initial value of weight correction factor (0.1)
Alpha W Min	Minimum value of weight correction factor (0)
Alpha W Inc	Decreasing amount of the weight correction factor (0.001)
Epochs	The epochs number for an experiment is automatically calculated by this formula: $Epochs = \frac{AlphaMax - AlphaMin}{AlphaInc}$

3. Input record

Patterns	Number of records in the input sample (144)

3.1. *Results*

Experiments performed with SOM gives the opportunity to analyze fuzzy similarities among the 144 Milan's zones.

As example, it is here illustrated the 5x5 units output matrix results.

Examining edge clusters profiles it becomes clear how the map layout is characterized by two main axes. Central city zones, the ones with a highest urban quality, are in the class 1-1, moving through the classes on the same matrix row a worsening of the social-economical features of the zones can be observed. The class 1-5 contains the suburban planned zones, where urban decay is present beside some advantages such as schools availability. In the other direction, through the column, social-economical quality remains constant, but an environmental decay can be observed, mainly referred to the building stock. In class 5-1 there are semi-central zones with high economic vitality, but with highly decayed buildings. Opposed to the central zones in class 1-1 there are the ones in class 5-5, semi-suburban, where quality is low for all the three considered aspects: social, economic and environmental; they are examples of urban segregation.

From a spatial point of view, the results are illustrated in the map in figure 3, where the Milan's spatial pattern as described by SOM can be represented.

The following descriptions are referred to the four analyzed clusters (fig. 6).

- **Profile 1-1**

  High values: Real estate values (particularly dwellings and offices), basic commercial services, home-work/study trips within 30 minutes, executives and businessmen, sq.m./inhabitant, housemaids.

  Medium values: crime, self employed, hotels.

The more central areas, inside the "cerchia dei navigli", belong to a well defined class, strongly characterized. It is clear underlying that the zones grouped together are the same in all the simulations, also changing the output matrix dimension. Distinctive features are the ones of central areas with relevant presence of business services. Here the urban quality reaches its maximum and the population is mainly composed by affluent and professional people.

- **Profile 1-5**

  High values: residential index, outgoing commuters, crowding index, automobile users rate, social services.

  Medium values: self employed, high school graduates, green areas.

Zones in this cluster are typical sub-urban areas marked by the presence of dormitory neighborhoods built during the urban expansion. These zones are mainly residential and their outgoing commuters rate is high. Nevertheless they are populated by middle classes: self-employed workers and people with high school degree. Usually in such zones there are green areas, being in the less dense suburbs, and a good amount of services, mainly crèches and nursery schools.

- **Profile 5-1**

  High values: crime, services to the population, pollution, building stock decay, entering commuters/surface,

  Medium values: residential index, outgoing commuters, outgoing commuters travelling within 30 minutes, offices prices.

Selected areas are semi-central and have strong presence of service industries, although not homogeneously distributed, and, as a consequence, high entering commuters rate and medium-high offices values. These areas seem to show a renewal dynamic related to the possible intervention to restore the relevant building stock decay.

- **Profile 5-5**

  High values: population prevailing age, youth unemployment, blue collars, public housing, outgoing commuters, pollution.

  Medium values: self employed, population density, social and base services to the population.

This cluster contains semi-suburban zones, with a particularly high decay degree, that are real social "ghettos". In this areas the population is composed by elderly, blue collars, and the young people is often unemployed. The residential index is very high, owing also to the relevant presence of public housing, and the population commutes outside the area.

The existence of indicators whose values are corresponding to the city mean shows that these areas, although unattractive and socially decayed, are integrated in the urban fabric. The population density is actually good, there are self-employed workers and also services for the people (shops, crèches, nursery schools and centres for the elderly).

Other interesting cases have been examined.

For instance, in cluster 3-5 (fig. 7) there are the zones where large derelict industrial areas are present (for example the Bovisa and Bicocca areas that are being transformed into university campus). Also areas with relevant presence of public green areas and low population density, corresponding with public parks, and finally suburban and semi-suburban zones where population settlement dynamic evolved in recent years, crime rate is low and travel time to reach the workplace is high.

4. Conclusions

SOM application enables the analysis of Milan’s urban system complexity, systematizing information given by social, economic and environmental indicators.

Classification problems are not new in the urban and territorial planning field, what is new is the use of an ANN method, an approach still largely ignored by the territorial analysis.

The question regarding if this kind of classification adds something respect more traditional statistical methods, such as *classification trees* or *factor analysis*, is improper. The theoretical and methodological premises are completely different. SOM use a fuzzy and self-organized approach, are unsupervised, not needing exogenous rules to be set, and can recognize not-linear correlation.

Relationships among units in the output map are based on fuzzy similarities and the unit pattern is meaningful on its own; preferential directions can be individuated like if on the map is superimposed a coordinates system.

The spatial approach typical of the SOM offers great perspective possibilities, it is a powerful tool to squeeze and synthesize information, able to build a base knowledge starting from a complex system.

Bibliography

BREDA M. (1999), Self- Organizing Maps, in Buscema M. & Semeion Group (ed.) *Reti Neurali Artificiali e Sistemi Sociali Complessi. Teoria e modelli*, v. I, Milan, Franco Angeli.

BUSCEMA M. & SEMEION GROUP (ed.) (1999), *Reti Neurali Artificiali e Sistemi Sociali Complessi. Teoria e modelli*, v. I, Milan, Franco Angeli.

BUSCEMA M. (2000), Reti Neurali e Problemi Sociali, Semeion Centro Ricerche, Working paper.

BUSCEMA M., DIAPPI L. (1999), La struttura complessa delle città, Reti Neurali per un sistema cognitivo, in Buscema M. (ed.) *Reti Neurali Artificiali e Sistemi Sociali Complessi. Applicazioni*, v. II, Milan, Franco Angeli.

DIAPPI L. (ed.) (2000), *Sostenibilità urbana*, Milan, Paravia.

DIAPPI L., BOLCHI P., CAMPEOL A., FRANZINI L. (1998), *La costruzione di una mappa di sostenibilità ambientale a Milano: le condizioni di rischio e opportunità* Ufficio Studi Camera di Commercio, Collana Argomenti e Ricerche - Strumenti per Milano.

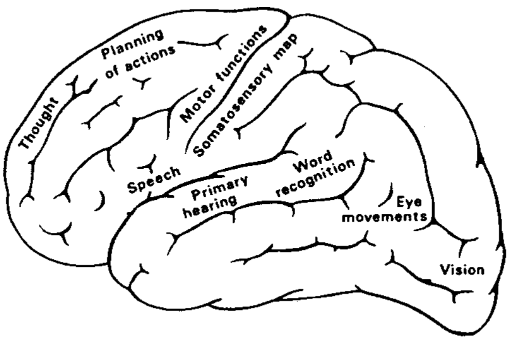
DIAPPI L., BOLCHI P., FRANZINI L. (1999), GIS e complessità urbana: un approccio neurale in Monti C., Besio M. (ed.), *Dal cannocchiale alle stelle*, Milan, Franco Angeli.

DIAPPI L., BUSCEMA M., OTTANÀ M. (1998), A neural Network investigation on the crucial Assets of urban Sustainability, *Substance Use and Misuse*, v. 33, pp. 793-817, ISBN/ISSN 1082-6084, Special Issue on Artificial Neural Networks and Social Systems .

DIAPPI L., FRANZINI L. (2000), Complessità urbana e misura del rischio, in Diappi L. (ed.) *Sostenibilità urbana*, Milan, Paravia.

KOHONEN T. (1995), *Self-Organising Maps*, Berlin, Springer Verlag.

PARISI D. (1989), *Intervista sulle reti neurali*, Bologna, Il Mulino.



A. Brain areas (Kohonen, 1995)

Figure 1 *Different brain areas connected to different sensorial modality*

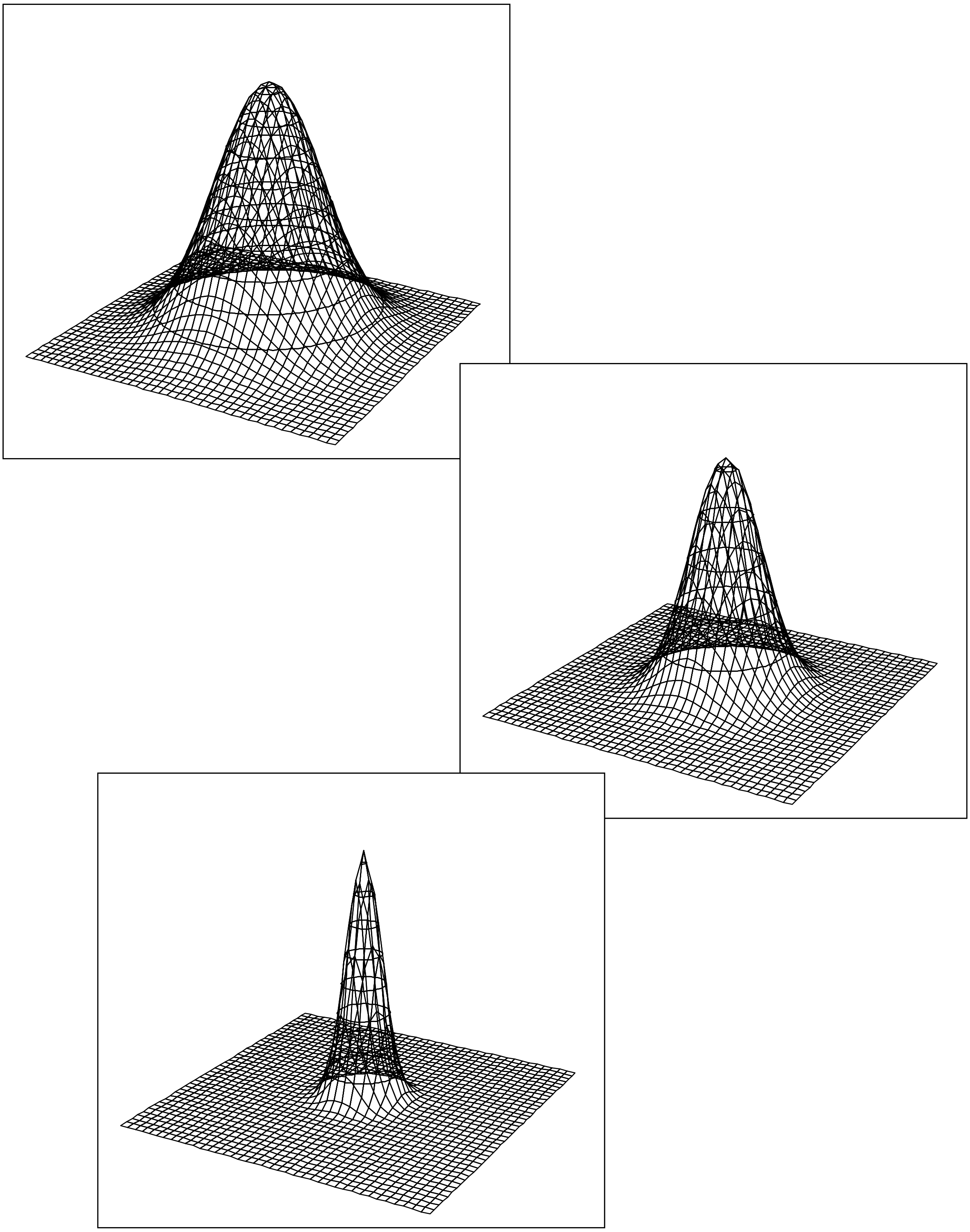


Figure 2 *Gaussian squeeze during the updating of the matrix.*

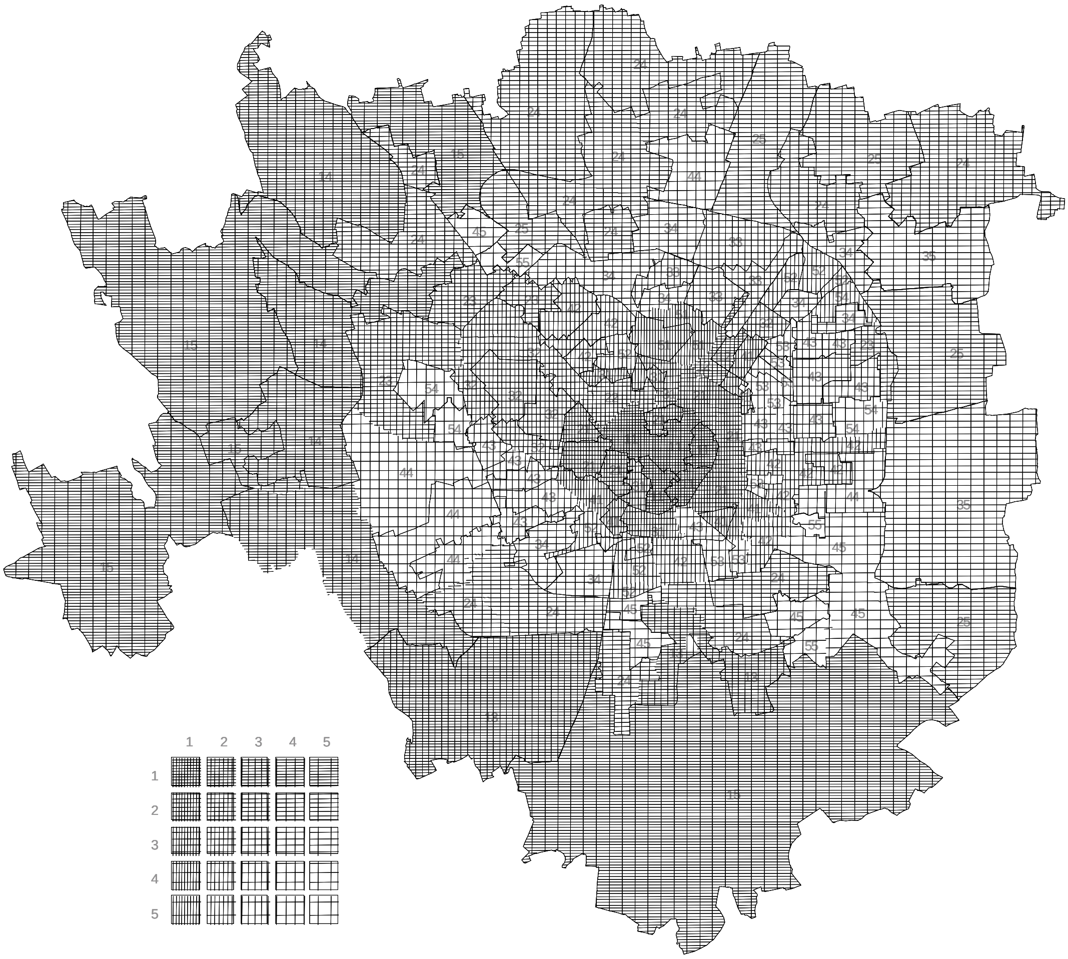


Figure 3 *Clustering of the 25 units matrix. Numbers and pattern help to link zones and units.*

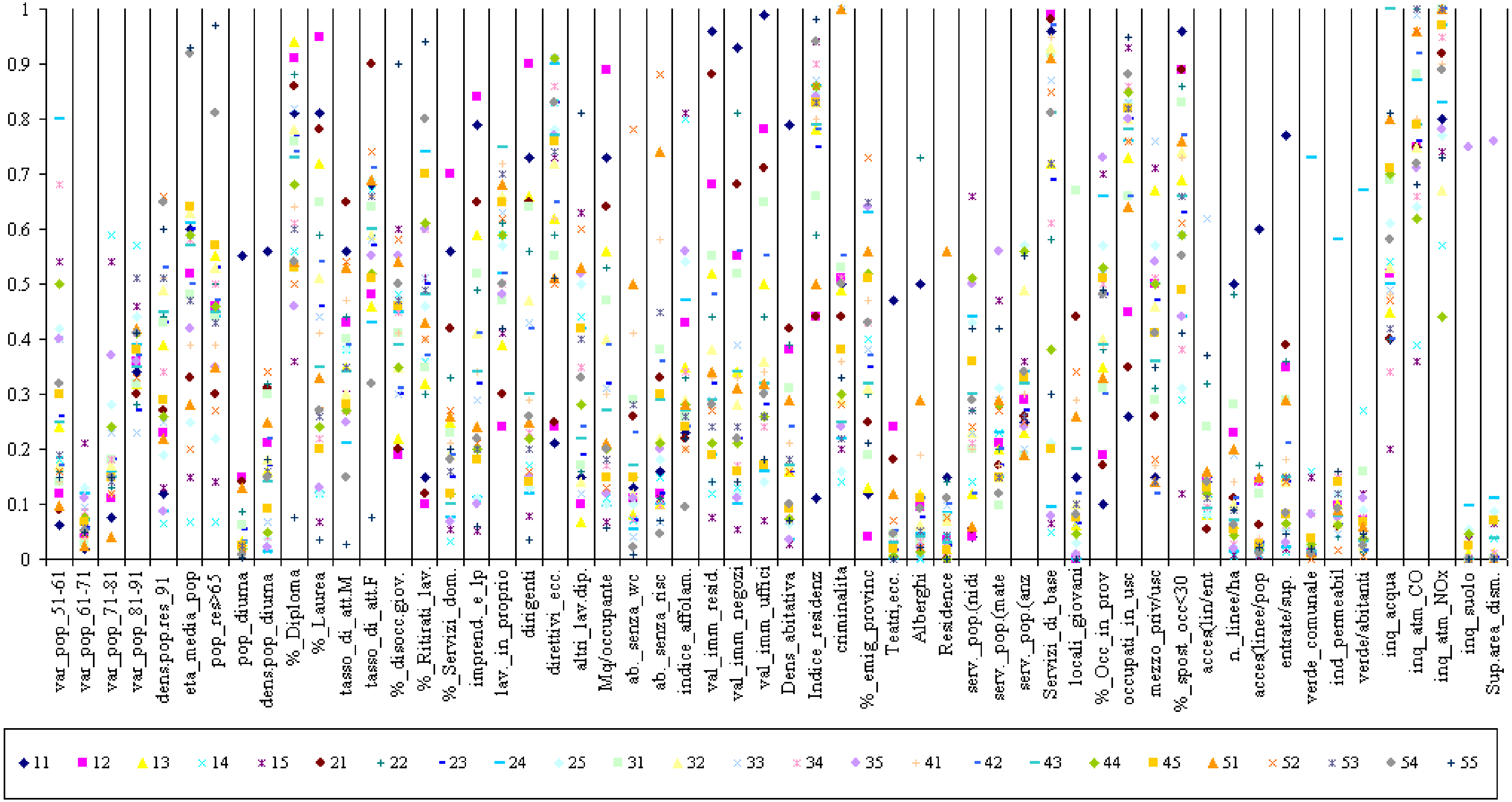


Figure 4 *Codebooks of the 25 units matrix. On the x-axis the 56 variables, on the y-axis the activation level of each codebook.*

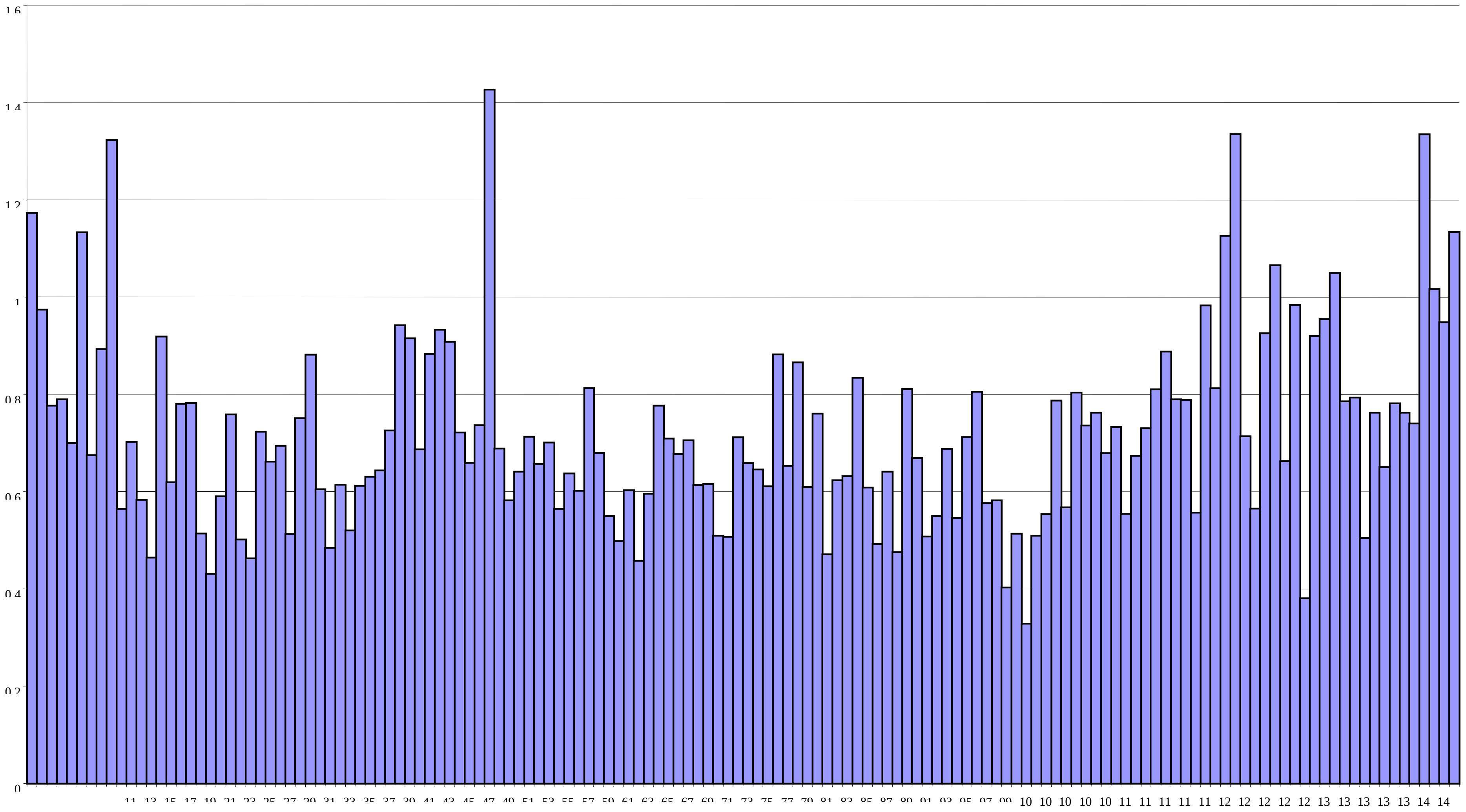


Figure 5 *Euclidean distances from hyper-points. On the x-axis the 144 zones, on the y-axis their distances from the codebook.*

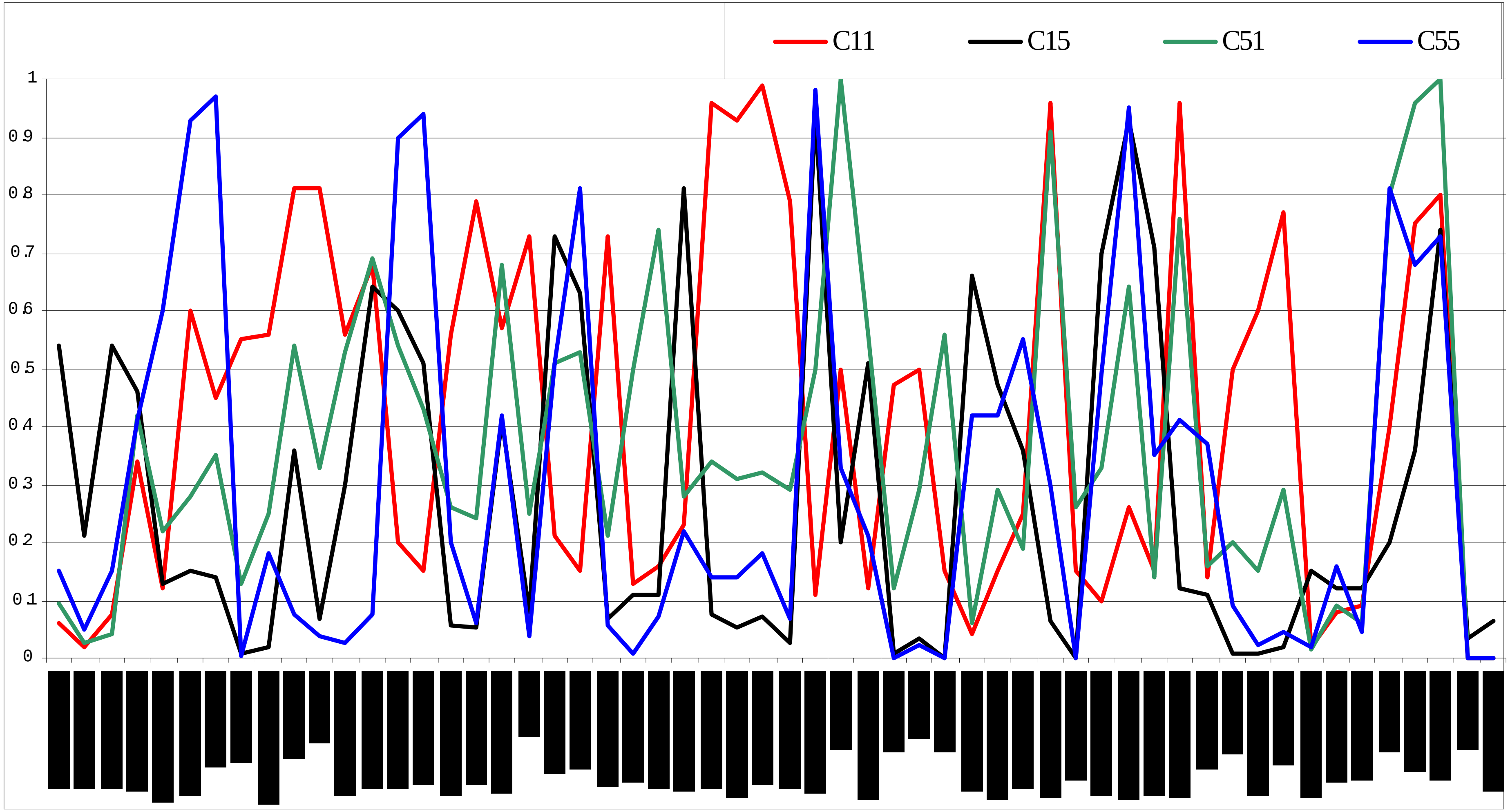


Figure 6 *Codebooks at the edges of the 25 units matrix.. On the x-axis the 56 variables, on the y-axis the activation level of the 4 codebooks.*

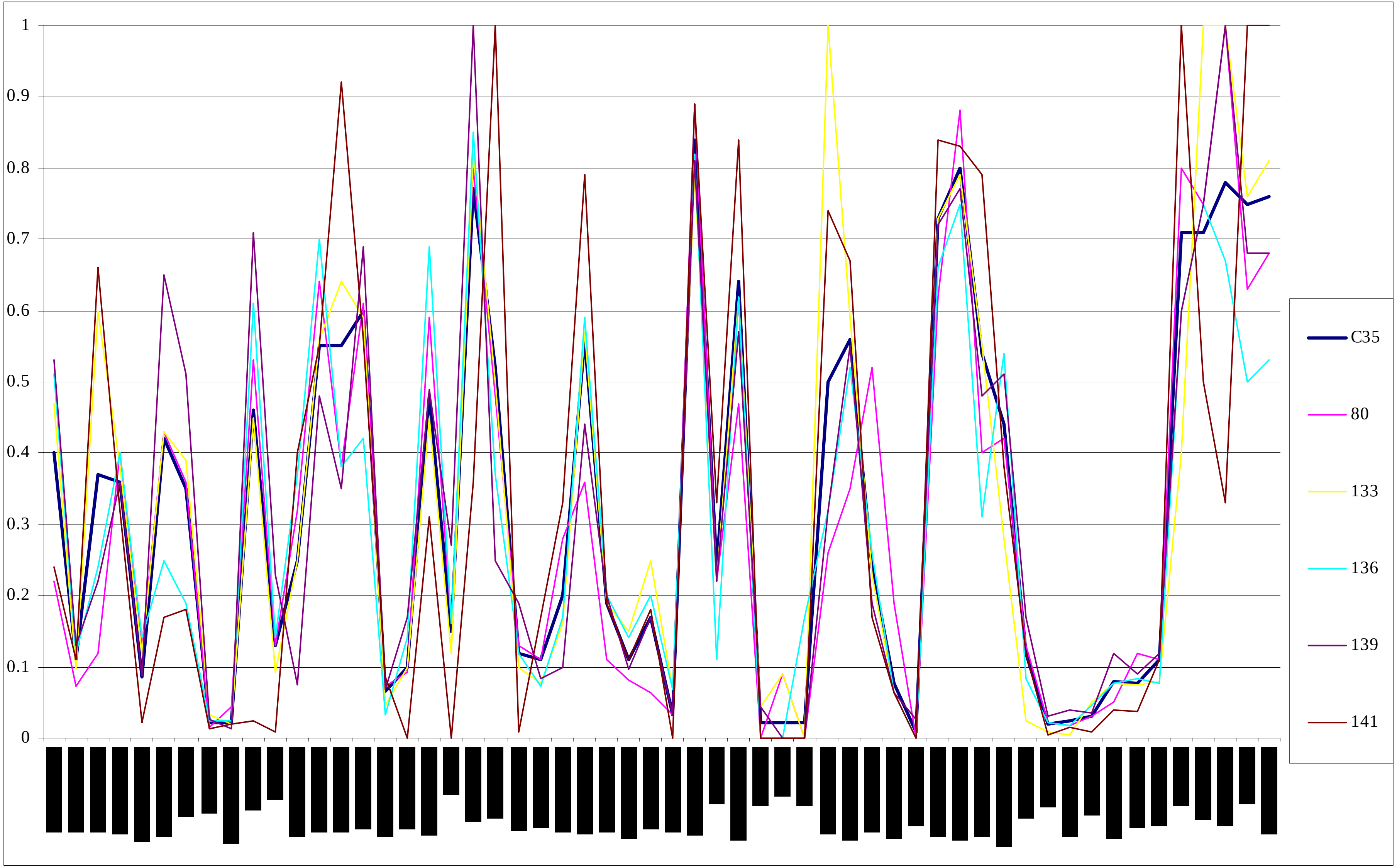


Figure 7 Example of c*luster profile of the 25 units matrix.. On the x-axis the 56 variables, on the y-axis the activation level of zones and codebooks.*