AN ULTRAMETRIC INTERPRETATION OF BUILDING RELATED EVENT DATA

PATRICK ERIK BRADLEY

Abstract. A random sample of event data from urban building stocks in Baden, Southwest Germany, is examined using ultrametric hierarchical classification.

1. Introduction

The building stock at a regional level is a highly complex entity, and studying its long-term dynamics is a most non-trivial undertaking. In addition, obtaining the necessary data, the crucial point of empirical studies, is often extremely difficult. A slight relief is the availability of earth scan data or digital cadastres. However, these provide information only for the last few years and are often still not publicly available in many places, and in any case costly.

For these reasons, much building-dynamic related research relies on official statistics, survival analyisis, random sampling, and theoretical models, or a combination of the four. To name only a few examples from the literature, the first method is probably ubiquitous (where available), the second has been used by [10, 11, 12, 13, 16], random sampling used in [6, 15], and the physical model approach in [5, 21].

The use of classification methods allows in principle the comparison of building stocks and their dynamics. Its use in the study of the urban built environment has already some tradition. In particular the dissertation [1] introduces not only the term *Urban Data Mining* for this research area, but also applies most recent data mining techiques to the building sector and combines these with the use of geographic information systems [2]. One possible gain can be creating knowledge through information transfer across a given class of building stocks.

The approach we pursue here is to compare the dynamics of different municipal building stocks through adopting an *ultrametric* point of view. The notion of ultrametricity refers to the geometry underlying hierarchical classification, more precisely the *dendrograms*, i.e. the tree-structure formed by the hierarchies in data. The objective of classification is often to find within data inherent hierarchical structure. The most natural geometric framework for this is provided by the ultrametric property of dendrograms: it is given by a distance function d on the dataset X satisfying the *strict triangle inequality*

$$d(x, y) \leq \max\{d(x, z), d(z, y)\} \tag{1}$$

for any x, y, z taken from dataset X. The function d measures the distance between x and y by computing the smallest cluster containing x and y, where clusters are defined by the nodes or levels of the dendrogram for X.

Date: April 14, 2009.

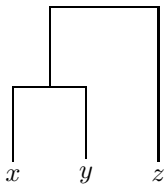

FIGURE 1. An ultrametric triangle.

The inequality (1) follows naturally from the tree-structure, as Figure 1 reveals: in that example, x and y form a strict subcluster of the dataset $X = \{x, y, z\}$, whereas any cluster containing z is either the singleton $\{z\}$ or the whole dataset X. Hence, x and y are ultrametrically closer to each other than to z.

Concerning the applications in hierarchical classification, the ultrametric distance allows the formulation of classification algorithms, often similar to their classical counterparts, and most often more efficient. The reason is that the dendrogram is uniquely determined by the ultrametric. Hence, in order to find the hierarchical structure underlying some given data X, the analyst must find a suitable ultrametric for X. As data (also categorical) can always be represented by rational numbers, the possible ultrametric distances are practically determined by the choice of a prime number. This fact is a theorem by Ostrowski [19]. Equivalently, one can fix a prime number p, and then the task is to find a suitable representation of data by rational or so-called *p-adic* numbers.

In the present case of this article, the data is a random sample of events on municipal building stocks taken from various archives. They allow the study of the event history of every sampled building, but this is not the focus here. The aim is to introduce a method for classifying such building stocks by applying ultrametric techiques to the dataset. As a result, we obtain a first approximation towards comparing certain dynamical aspects of some urban building stocks in the region of Baden in Southwest Germany.

In the following second section, we briefly review the methods from ultrametric hierarchical classification used in this article. The third section is a short introduction to Fourier analysis. The reason is that the Fourier transform takes multidimensional data to the one-dimensional frequency domain. As a consequence, the linear ordering of frequencies allows, after quantisation, a direct binary encoding of the transformed data and use of the 2-adic distance. Section 4 provides more details on the dataset and its acquisition. In Section 5 a segmentation of the time line through ultrametric classification of decades is performed, i.e. time quanta are classified. The methods used are the Fourier transform and a topology-driven alternative. Section 6 contains a comparison of the results from the previous section and a direct Fourier transform approach. The final Section concludes the article and is followed by the appendix containing the bulk of the figures used in the article.

2. Ultrametric hierarchical classification: a primer

As the dendrograms in this article are all binary, we discuss only that case in this introductory section. For more details and also the general case, we refer to [8, 9].

2.1. **Labeled dendrograms.** An important object in hierarchical classification is the *dendrogram*, a tree-like representation of hierarchies within data. Figure 2 shows a dendrogram generated from the dataset used in this article.

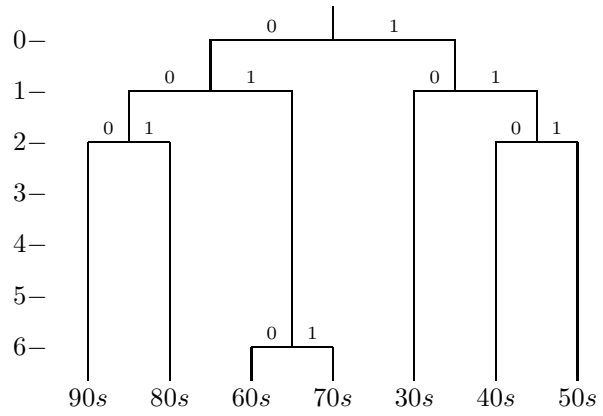

FIGURE 2. Dendrogram for decades from whole dataset.

At the bottom end of the dendrogram lies the data, and each horizontal line segment represents a cluster of those data at the end of the downward paths from that line segment. The horizontal lines in the dendrogram are labeled '0' or '1' in each level. This allows to encode the data with binary numbers in the following way. A path from the top down to some datum x will 'pick up' label $a_\nu \in \{0, 1\}$ at each level ν. Then the binary number $B(x)$ is

$$B(x) = \sum_\nu a_\nu 2^\nu,$$

where ν runs through the levels. This method allows to use the 2-adic distance on the dataset X:

$$d(x, y) = 2^{-\mu(x,y)},$$

where

$$\mu(x, y) = \max\{\nu \mid 2^\nu \text{ divides } B(x) - B(y)\},$$

the lowest exponent appearing in $B(x) - B(y)$ viewed as a sum of powers of 2. An immediate consequence is that the more initial terms $B(x)$ and $B(y)$ have in common, the smaller the distance between x and y will be.

The 2-adic distance is an example of a so-called *ultrametric*, and satisfies its characteristic property:

$$d(x, y) \leq \max\{d(x, z), d(z, y)\}$$

for any $z \in X$. We will also speak of d as a *p-adic metric*, where p is a prime number (in our case, $p = 2$). The general choice of p allows for the non-binary case by labelling the n branches at a given level with $0, 1, \ldots, p - 1$, if $p \geq n$. However, we will not explain why p is often perferred to be a prime number.

2.2. **p-adic classification.** Hierarchical classification of a dataset X usually depends on the choice of a metric. There is a vast literature for the case that data is represented by vectors in a Euclidean space $\mathbb{R}^n$, together with the Euclidean metric.

Using the p-adic metric implies many simplifications due to the fact that the dendrogram is uniquely determined by the binary (or, more generally, p-adic) representation of X. A consequence is that hierarchical agglomerative or chain clustering algorithms perform much faster than their classical counterparts [3, 8]. One disadvantage of this approach is that the data needs a binary (or p-adic) encoding. There is usally no canonical way of doing this. In Sections 5 and 6, data-driven approaches to this end are described.

The classification algorithm we use with the dataset of this article minimises an energy function

$$E(\mathcal{C}, k) = \sum_{C \in \mathcal{C}} (\#C - 1) \cdot \mu(C)$$

where $\#C$ denotes the cardinality of C, $\mu(C) = \max\{d(x, y) \mid x, y \in C\}$, and $\mathcal{C}$ is a clustering of X containg at most k clusters. In [9], this approach is described in detail. The upper bound k for the cluster number has to be prescribed in advance. We choose it as the point of maximal curvature of an exponential fit to the function

$$L(k) = \frac{1}{k} \sum_{C \in \mathcal{C}} \mu(C).$$

This is in analogy to the classical case, where such an approach is believed to yield optimal cluster numbers.

3. Fourier transform of time series

One important tool in signal processing is the discrete Fourier transformation (DFT). Given a real or complex periodic signal f, the Fourier transform decomposes it into the different frequency parts "contained" in it by writing it as a sum of trigonometric functions. The idea is to consider the signal as a superposition of waves of different wavelength and amplitude. In the discrete case, $f = (f_0, \dots, f_{N-1})$ is a vector of finite dimension, and the Fourier transform $\hat{f}$ is given as

$$\hat{f}_k = \sum_{j=0}^{N-1} f_j e^{-2\pi i \frac{jk}{N}}, \quad k = 0, \dots, N-1,$$

where $\hat{f}_k$ is the k-th component of the transformed vector $\hat{f}$, and $e^{-2\pi i x}$ is the complex exponential function.

The Fourier transform takes a signal from the time domain to the frequency domain, and the absolute value $|\hat{f}_k|$ is the *amplitude* of $\hat{f}$.

For visualising the Fourier transform of a signal, the amplitude is often plotted against frequency k. However, especially in the presence of noise it is often convenient to suppress it in the representation or at least decrease its visibility. This can be done for example by further logarithmic transformation:

$$\hat{f} \to \log |\hat{f}|$$

often has this desired property.

In the case of a real-valued signal, the Fourier transform has the property

$$\hat{f}_{N-k} = \hat{f}_k,$$

implying that there are only $\frac{N}{2}$ independent Fourier coefficients. This explains the symmetry of e.g. Figures 10 and 11.

4. The event data

The dataset used in this article is a random sample taken from building insurance files which record the history of insurance-relevant events occuring to insured buildings in the former state of Baden (Germany) between 1936 and 1993. This interval is the time of insurance monopoly and legal obligation. More details can be found in [6].

Name	code	# lots	# buildings	# events
Event dataset	ALL	95	279	896
Baden-Baden	BAD	1	11	51
Freiburg	FR	19	68	231
Heidelberg	HD	11	24	79
Mannheim	MA	23	69	186
Pforzheim	PF	31	72	228
Rastatt	RA	3	18	73
Eberbach, Eppingen, Gaggenau, Neuhausen, Villingen	other	7	17	48

Table 1. The event dataset.

The files of addresses containing insured buildings were kept in the corresponding municipality at the time, and were transferred to an archive during 1994. The "dead" files correspond to addresses whose buildings were no longer insured under that address, e.g. because of demolition or property splitting. The insurance retains a list of those addresses and left it to the decision of each municipality to either keep the dead files or transfer it to the archive. With a small number of exceptions, the dead files were retained locally. We call this list the *dead-file list*.

Identifier	Meaning
c	number of contour changes
d	number of demolitions
f	number of function changes
n	number of new constructions
o	number of owner changes
r	number of renovations
s	number of changes in storey number

Table 2. The variables in event data set.

In the present case, a random sample of 100 addresses was taken from the dead-file list, and the corresponding files were extracted first from the archive and requested from the municipalities. Not all archives were able to provide copies of the requested files, and some included files on unrequested addresses. The dataset now consists of information on 279 buildings from 95 addresses. Table 1 shows the numbers of addresses, buildings and events found in each municipality before the pre-processing of data. This means that war-destruction is counted as an event, and some coincident events events are not counted with the correct multiplicity. Table 2 lists the kind of information extracted for the actual dataset of this article, and in Figure 5 the counts of the different event types are depicted.

A first observation in Figure 5 is that, although war-destruction was eliminated from the dataset, there do exist war-related peaks in the 1940s and 1950s. Somewhat surprising might be that the number of new constructions seems rather low in the post-war era, and especially after 1960. This is due to the special property of the dataset as being sampled from the dead files. In the following sections, we aim at understanding the dynamics within the dataset from an ultrametric point of view.

As a first processing step, the variables were quantized by setting each value 0 or 1, depending on whether the corresponding count is low or high. The procedure applied for finding the border line between 0 and 1 is quite similar to the finding of optimal cluster numbers. Let $v\colon T \to \mathbb{N}$ be an event variable, i.e. counting the number of events at time $t \in T = \{t_1, \ldots, t_n\}$. Then we denote by

$$b(\tau) := \#\{t \in T \mid v(t) > \tau\}$$

the *cut function* in $\tau \in \mathbb{R}$. Again the point of highest curvature in an exponential fit to $b(\tau)$ determines the border line for quantisation.

5. Time segmentation by hierarchy

In order to obtain a view on the dynamics of the sampled building stock, the timeline 1936-1993 is partitioned into the decades 30s to 90s. The aim is to classify the decades in order to obtain a segmentation of the time line. The method used for this end is ultrametric hierarchical classification.

5.1. **Segmentation by Fourier transformation.** The idea of ultrametric or p-adic classification is to use an ultrametric distance on data in order to obtain a unique dendrogram and then to estimate clusters. This is most conveniently realised if the data is encoded e.g. with binary numbers (more generally, one can use so-called *p-adic numbers*). In order to obtain a binary encoding of data, one can follow the authors of [3] by using a Fourier transform and work in the frequency domain. There, the ranking of frequencies is simply by magnitude. The low frequencies are given the highest ranks in the hierarchy, and higher frequencies become less important.

In order to find a ranking of the variables c, d, f, n, o, r, s from Table 2, an adaptation of the above the idea would be to transform the data and then compare the distributions in the frequency domain. A variable with more support in the lower frequency-range is then ranked higher than a variable with comparatively smaller low-frequency amplitudes. The discrete Fourier transforms of the aggregated counts

by decades are plotted on a logarithmic scale in Figure 10. The corresponding amplitudes for the lowest frequency yield the ordered sequence

$$n,\ o,\ s,\ r,\ d,\ f,\ c,$$

which is confirmed by the Fourier transforms of the aggregated event count by five-year periods (cf. Figure 10).

	$30s$	$40s$	$50s$	$60s$	$70s$	$80s$	$90s$
n	1	1	1	0	0	0	0
o	0	0	1	1	1	0	0
s	0	1	1	1	1	0	0
r	0	1	1	1	1	1	0
d	0	0	1	1	1	1	0
f	0	1	1	0	1	0	0
c	0	1	1	1	1	0	0

TABLE 3. Quantised counts (decades, whole dataset).

The aggregated values for the different decades are quantised as in Table 3. The corresponding decade-histograms are depicted in Figure 6. The quantisation of the counts into low = 0 and high = 1 was obtained as explained in the end of Section 4. From the quantised values, the dendrograms can be read off immediately. They are shown in Figure 8. The optimal clustering method then yields the following segmentations:

$$\begin{aligned}
\text{ALL} &\quad |\ \underbrace{30s\ 40s\ 50s}_{A}\ |\ \underbrace{60s\ 70s}_{B_1}\ |\ \underbrace{80s}_{B_2}\ |\ \underbrace{90s}_{B_1}\ | \\
\text{FR} &\quad |\ \underbrace{30s\ 40s\ 50s}_{A}\ |\ \underbrace{60s\ 70s}_{B_1}\ |\ \underbrace{80s}_{B_2}\ |\ \underbrace{90s}_{B_1}\ | \\
\text{MA} &\quad |\ \underbrace{30s\ 40s\ 50s}_{A}\ |\ \underbrace{60s\ 70s\ 80s\ 90s}_{B}\ | \\
\text{PF} &\quad |\ \underbrace{30s\ 40s\ 50s}_{A}\ |\ \underbrace{60s\ 70s}_{B}\ |\ \underbrace{80s}_{A}\ |\ \underbrace{90s}_{B}\ | \\
\text{HD} &\quad |\ \underbrace{30s\ 40s}_{A_2}\ |\ \underbrace{50s}_{A_1}\ |\ \underbrace{60s}_{B}\ |\ \underbrace{70s}_{A_1}\ |\ \underbrace{80s\ 90s}_{B}\ | \\
\text{RA} &\quad |\ \underbrace{30s\ 40s\ 50s}_{B}\ |\ \underbrace{60s}_{A}\ |\ \underbrace{70s}_{B}\ |\ \underbrace{80s}_{A}\ |\ \underbrace{90s}_{B}\ | \\
\text{BAD} &\quad |\ \underbrace{30s}_{B_2}\ |\ \underbrace{40s}_{A}\ |\ \underbrace{50s\ 60s\ 70s}_{B_1}\ |\ \underbrace{80s\ 90s}_{B_2}\ | \\
\text{other} &\quad |\ \underbrace{30s}_{A}\ |\ \underbrace{40s}_{B}\ |\ \underbrace{50s\ 60s\ 70s}_{A}\ |\ \underbrace{80s}_{B}\ |\ \underbrace{90s}_{A}\ |
\end{aligned}$$

Observe that the period 1930-1960 was governed by new construction in the larger urban building stocks ALL, FR, MA, PF, HD, whereas in the smaller ones this was only partially the case (BAD, other) or not at all (RA).

5.2. **Topological segmentation.** An alternative approach proposed here is *ranking by topology*. For this, the first ranking criterion is now given by the number of contiguous sequences of $1, \ldots, 1$ and $0, \ldots, 0$. This puts n to the top and f to the bottom. In order to break the ties among the other variables, we apply the same criterion to the quantised five-year aggregated counts in Table 4. This puts c higher than d and o, which in turn are higher ranked than r and s. Note that the topological ranking for the five year counts does not contradict the one for the decade counts. The remaining ties d, o and r, s are now broken by ranking higher in both cases the variable with longer contiguous $1, \ldots, 1$ sequence in the corresponding decade count. The ranking obtained in this way is n, c, d, o, r, s, f. Note that doing the same with the five-year counts would yield the different ranking n, c, o, d, s, r, f. However, since the main concern is a classification of decades, it seems more appropriate to give higher priority to the decade counts, and use the five year counts only for tie-breaking purposes.

	30.2	40.1	40.2	50.1	50.2	60.1	60.2	70.1	70.2	80.1	80.2	90.1
n	1	1	1	1	1	0	0	0	0	0	0	0
c	0	0	1	1	1	1	1	1	1	0	0	0
d	0	0	0	0	1	1	1	1	1	1	0	1
o	0	0	1	1	1	1	1	1	1	1	0	1
r	0	1	1	1	1	0	1	1	1	0	0	0
s	0	0	1	1	1	1	1	0	1	1	0	0
f	0	0	1	1	1	0	1	0	1	0	1	0

TABLE 4. Quantised counts (five-year blocks, whole dataset).

From Table 3, after re-ordering the rows according to the new ranking, it is now possible to read off the dendrogram. It is the one depicted in Figure 2 from Section 2. The method for the optimal number of clusters yields 3 in this case, and the segmentation of the time line is

$$| \underbrace{30s\ 40s\ 50s}_{A} | \underbrace{60s\ 70s}_{B_1} | \underbrace{80s\ 90s}_{B_2} |$$

with a major change point in the dynamics around 1960. The municipality-wise computed dendrograms are depicted in Figure 9 in Appendix A. This yields the

segmentations:

$$
\begin{array}{ll}
\text{ALL} & |\ \underbrace{30s\ 40s\ 50s}_{A}\ |\ \underbrace{60s\ 70s}_{B_1}\ |\ \underbrace{80s\ 90s}_{B_2}\ | \\
\text{FR} & |\ \underbrace{30s\ 40s\ 50s}_{A}\ |\ \underbrace{60s\ 70s}_{B_1}\ |\ \underbrace{80s\ 90s}_{B_2}\ | \\
\text{MA} & |\ \underbrace{30s}_{A_1}\ |\ \underbrace{40s\ 50s}_{A_2}\ |\ \underbrace{60s\ 70s\ 80s\ 90s}_{B}\ | \\
\text{PF} & |\ \underbrace{30s}_{A_1}\ |\ \underbrace{40s\ 50s}_{A_2}\ |\ \underbrace{60s\ 70s}_{B}\ |\ \underbrace{80s}_{A_1}\ |\ \underbrace{90s}_{B}\ | \\
\text{HD} & |\ \underbrace{30s}_{A_1}\ |\ \underbrace{40s\ 50s}_{A_2}\ |\ \underbrace{60s}_{B}\ |\ \underbrace{70s}_{A_1}\ |\ \underbrace{80s\ 90s}_{B}\ | \\
\text{RA} & |\ \underbrace{30s\ 40s\ 50s}_{B}\ |\ \underbrace{60s}_{A_1}\ |\ \underbrace{70s}_{B}\ |\ \underbrace{80s}_{A_2}\ |\ \underbrace{90s}_{B}\ | \\
\text{BAD} & |\ \underbrace{30s}_{B_1}\ |\ \underbrace{40s}_{A}\ |\ \underbrace{50s}_{B_2}\ |\ \underbrace{60s\ 70s\ 80s\ 90s}_{B_1}\ | \\
\text{other} & |\ \underbrace{30s}_{A_1}\ |\ \underbrace{40s}_{B}\ |\ \underbrace{50s\ 60s}_{A_1}\ |\ \underbrace{70s}_{A_2}\ |\ \underbrace{80s}_{B}\ |\ \underbrace{90s}_{A_2}\ |
\end{array}
$$

Notice that, from Tables 3 or 4, it may seem surprising that the main period of new construction seems to be before 1960, and not the 60s and 70s. This is, of course, due to the special nature of the dataset itself as coming from information on addresses whose records "dropped" out between 1936 and 1994.

6. Comparing dynamics of urban building stocks

In this section, we use the segmentations from the previous section and a direct Fourier transformation method in order to compare the dynamics of the different urban building stocks in our dataset.

6.1. **Dynamics from segmentation.** Comparing the segmentations in the previous section yields as a first observation that the two approaches DFT and topology yield similar results. There is a group ALL, FR, PF, MA, HD for which the first three decades are characterised by a high new construction activity. The other group RA, BAD, other has later or more periods in which new construction is dominant. The first group can be subdivided into ALL, FR, MA for which the 60s to early 90s are concerned with more emphasis on refurbishment related activities, and into HD, PF with another new construction period in the 70s or 80s, respectively.

This yields for both methods the clustering of the overall dynamics:

$$\text{ALL, FR, MA} \mid \text{HD, PF} \parallel \text{BAD, RA, other} \tag{2}$$

and compares well to the findings of [1], where it was observed that the larger municipalities are follow different dynamics from the smaller ones.

6.2. **A direct Fourier transform approach.** The Fourier transformation approach can be also applied to the sequence of all event counts for each municipality. Table 5 presents the normalised Fourier coefficients, where normalisation was obtained through division by the coefficients corresponding to ALL. Quantisation was performed with break at 0.7. This was chosen in order to have a distinction already

Coeff.	ALL	BAD	FR	HD	MA	PF	RA	other
0	1	0.60	0.81	0.65	0.78	0.79	0.62	0.62
1	1	0.55	0.78	0.66	0.78	0.74	0.64	0.61
2	1	0.61	0.79	0.69	0.82	0.80	0.75	0.73
3	1	0.71	0.68	0.42	0.50	0.76	0.58	0.77
4	1	0.71	0.96	0.28	0.79	0.27	0.56	0.66
5	1	0.35	0.96	0.43	0.78	0.45	0.43	0.49
6	1	0.71	0.96	0.28	0.74	0.27	0.56	0.65
7	1	0.71	0.68	0.42	0.50	0.77	0.58	0.77
8	1	0.61	0.79	0.69	0.82	0.80	0.75	0.73
9	1	0.55	0.78	0.66	0.78	0.74	0.64	0.61

TABLE 5. FFT of events.

at the lowest frequency (first row of Table 5), which can also be achieved with the ellbow method applied before. The corresponding dendrogram is depicted in Figure 3, and we obtain two optimal clusters

(3) ALL, FR, PF, MA, || HD, BAD, RA, other

for the overall dynamics. The clusterings (2) and (3) are quite similar, with the exception of a regrouping of HD. The Fourier transformation applied to the aggregated five-year counts yields the same classification, as can be verified from Figure 3. The main effect is that now FR and MA can be distinguished.

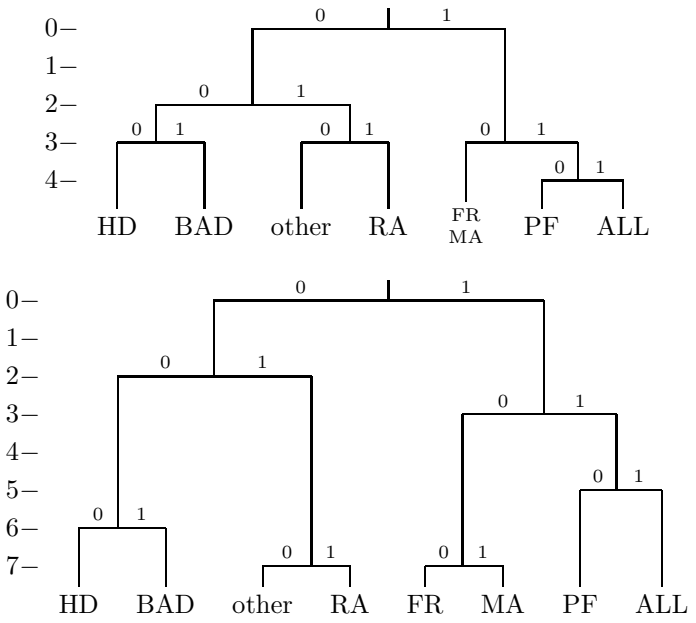

FIGURE 3. Municipalities dendrogram from FFT of events. Upper: decade-wise, lower: five-year-wise.

6.3. **Comparing event types.** In order to understand the dynamics of the different event types relative to each other, we apply the method of the previous subsection to the variables c, d, f, n, o, r, s representing the different types of event in the dataset.

A first observation in Figure 10 is that for all event types, the low frequency components are at high amplitudes, hence distinction can be made only in the middle or high frequency domain of the spectrum. Notice that the high frequency range appears in the middle of the diagrams in Figure 10, because of the cyclic property of the discrete Fourier transformation. This holds true for the aggregated counts by decades as well as by five-year periods, as can be checked by comparing Figures 10 and 11.

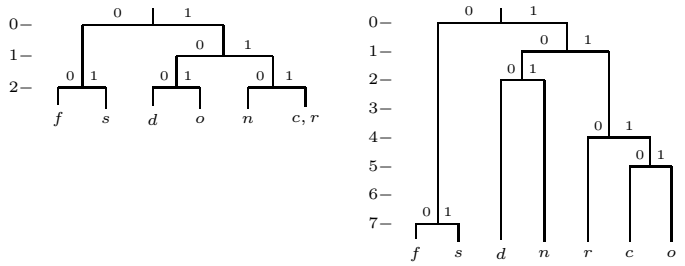

FIGURE 4. Dendrograms of events after DFT. Left: decades, right: five-year periods.

The findings after amplitude quantisation are the dendrograms in Figure 4. The aggregated decade counts yield the optimal clustering

$$f,\ s \,||\, d,\ o \,|\, c,\ n,\ r,$$

and the aggregated five-year counts yield:

$$f,\ s \,|\, c,\ d,\ n,\ o,\ r,$$

confirming the divide into the two event types f, s with small amplitudes in the middle-frequency regions of the spectrum, and the other types with high amplitudes in that same frequency domain.

7. Conclusion

The ultrametric approach allows the comparison of the dynamical behaviour of building stocks from a bird's eye view. The dynamics are described here by a segmentation of the time line into decades of similar activity. The Fourier transformation takes multi-dimensional data to the one-dimensional frequency domain. The lowest-frequency values allow a ranking of the event variables in order to obtain a binary data encoding on which ultrametric classification algorithms can be applied. A topology-based ranking of the event variables provides an alternative binary encoding scheme, and yields a similar segmentation of the time line as the Fourier transform approach. The linear ordering of frequencies yields a further binary encoding of Fourier-transformed data in order to obtain a more direct classification of municipalities with respect to their building dynamics.

The dataset used in this article is a random sample taken from the set of addresses in Baden, Southwest Germany, whose insurance files terminate between 1936 and 1993. The ultrametric approach to these data reveals a similar behaviour among the larger urban building stocks in contrast to the smaller ones. This is in correspondence to the findings from official statistics data.

Acknowledgements

The author acknowledges support from the DFG-project BR 3513/1-1. The lists of addresses from which the data was retrieved were generously entrusted by H. Gerstner from Sparkassenversicherungen SV, and the data itself was provided by its archive and the archives of Baden-Baden, Eberbach, Eppingen, Freiburg, Gaggenau, Heidelberg, Landratsamt Enzkreis, Mannheim, Pforzheim, Rastatt and Villingen-Schwenningen. Thanks to Martin Behnisch and Boris Jutzi for many valuable discussions. The Institut für Photogrammetrie und Fernerkundung at University Karlsruhe is warmly thanked for the opportunity to write down this article.

References

[1] Martin Behnisch. *Urban Data Mining. Operationalisierung und Strukturbildung von Ähnlichkeitsmustern über die gebaute Umwelt.* Dissertation. Universitätsverlag Karlsruhe (2008)

[2] Martin Behnisch and Alfred Ultsch. *Urban Data Mining Using Emergent SOM.* In: C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker (eds.) *Data Analysis, Machine Learning and Applications.* Springer, Berlin (2008)

[3] J. Benois-Pineau, A.Yu. Khrennikov, N.V. Kotovich. *Segmentation of Images in p-Adic and Euclidean Metrics.* Dokl. Math., 64, 450–455 (2001)

[4] J. Benois-Pineau and A.Yu. Khrennikov. *Significance Delta Reasoning with p-Adic Neural Networks: Application to Shot Change Detection in Video.* The Computer Journal. In Press. DOI: 10.1093/comjnl/bxm087.

[5] H. Bergsdal, H. Brattebø, R.A. Bohne and D.B. Müller. *Dynamic flow analysis for Norway's dwelling stock.* Building Research & Information, 45, 557–570 (2007)

[6] Patrick Erik Bradley and Niklaus Kohler. *Methodology for the survival analysis of urban building stocks.* Building Research & Information, Vol. 35, Nr. 5, 529-542 (2007)

[7] Patrick Erik Bradley. *Degenerating Families of Dendrograms.* J. Classif., 25, 27–42 (2008)

[8] Patrick Erik Bradley. *Mumford dendrograms.* The Computer Journal. In Press. DOI: 10.1093/comjnl/bxm088.

[9] Patrick Erik Bradley. *On p-adic Classification.* Preprint arXiv:0903.2870v1 [cs.AI]

[10] M.E. Gleeson. *Estimating housing mortality from loss records.* Environment and Planning A, 17, 647–659 (1985)

[11] M.E. Gleeson. *Estimating housing mortality with standard loss curves.* Environment and Planning A, 18, 1521-1530 (1986)

[12] I.M. Johnstone. *Energy and mass flows of housing. Estimating mortality.* Building and Environment, 36, 43–51 (2001)

[13] I.M. Johnstone. *Energy and mass flows of housing.* A model and example. Building and Environment, 36, 27–41 (2001)

[14] Andrei Khrennikov and Brunello Tirozzi. *Algorithm of Learning of p-adic Neural Networks.* Preprint.

[15] N. Kohler, U. Hassler and H. Paschen (eds.). *Stoffströme und Kosten im Bereich Bauen und Wohnen. Studie im Auftrag der Enquete-Kommission des deutschen Bundestags 'Schutz des Menschen und der Umwelt'.* Springer, Berlin (1999)

[16] G. Meinen, P. Verbiest and P.-P. de Wolf. *Perpetual Inventory Method. Service lives, discard patterns and depreciation methods.* Statistics Netherlands, Department of National Accounts (1998)

[17] Fionn Murtagh. *On ultrametricity, data coding, and computation.* Journal of Classification, 21, 167–184 (2004)

[18] Fionn Murtagh. *From Data to the p-Adic or Ultrametric Model.* p-Adic Numbers, Ultrametric Analysis and Applications, 1, 53–63 (2009)

[19] Alexander Ostrowski. *Über einige Lösungen der Funktionalgleichung* $\varphi(x).\varphi(y) = \varphi(xy)$. Acta math., 41, 271–284 (1916)

[20] Christophe Perruchet. *Hierarchical Classification of Mathematical Structures.* Stat. Prob. Lett., 1, 61–67 (1982)

[21] I. Sartori, H. Bergsdal, D.B. Müller and H. Brattebø. *Towards modelling of contsruction, renovation and demolition activities: Norway's dwelling stock, 1900–2100.* Building Research & Information, 36, 412–425 (2008)

Appendix A. Figures

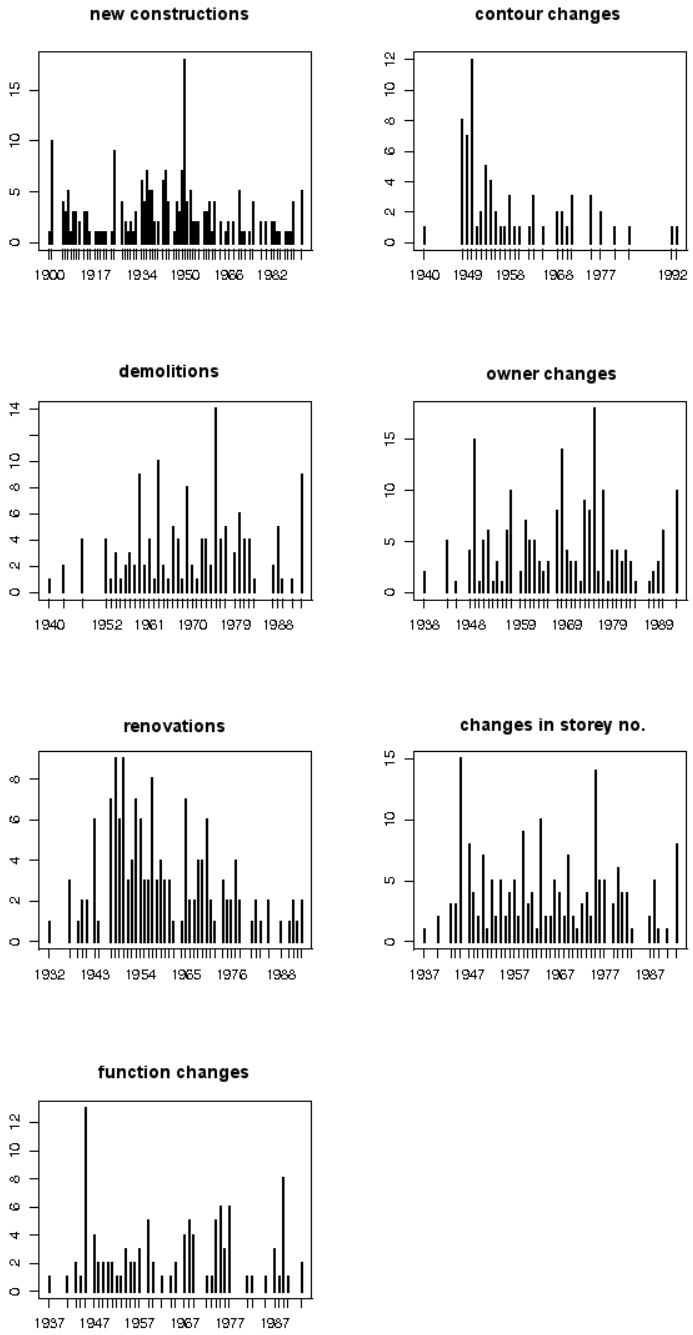

Figure 5. Event counts (event type).

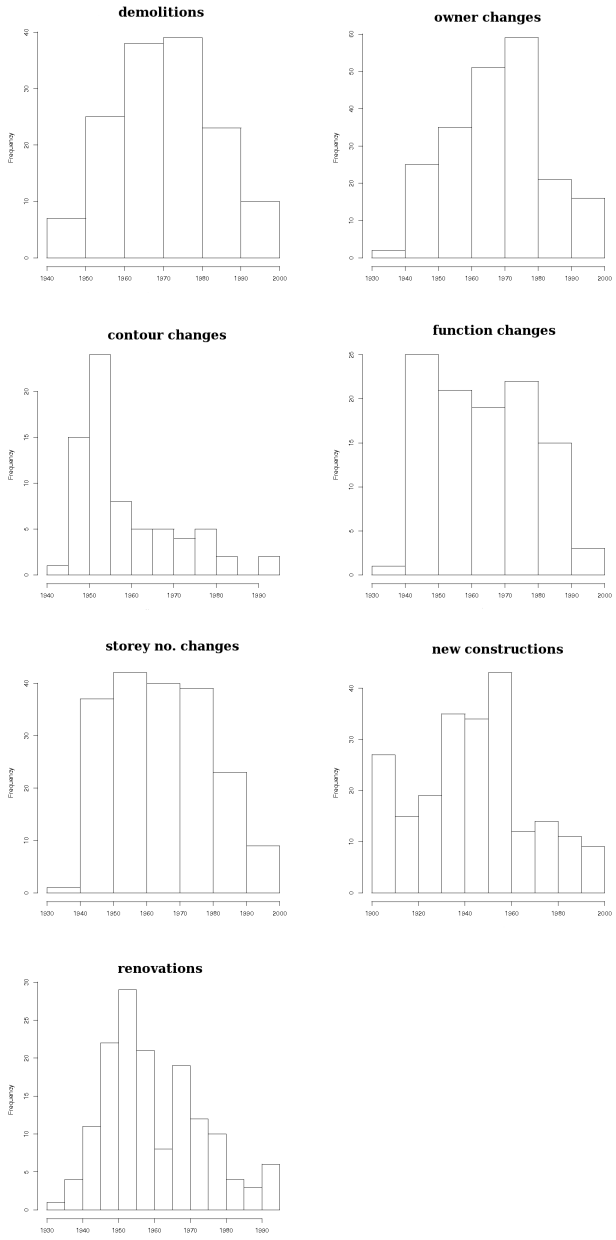

FIGURE 6. Decade-historgrams of event counts (event type).

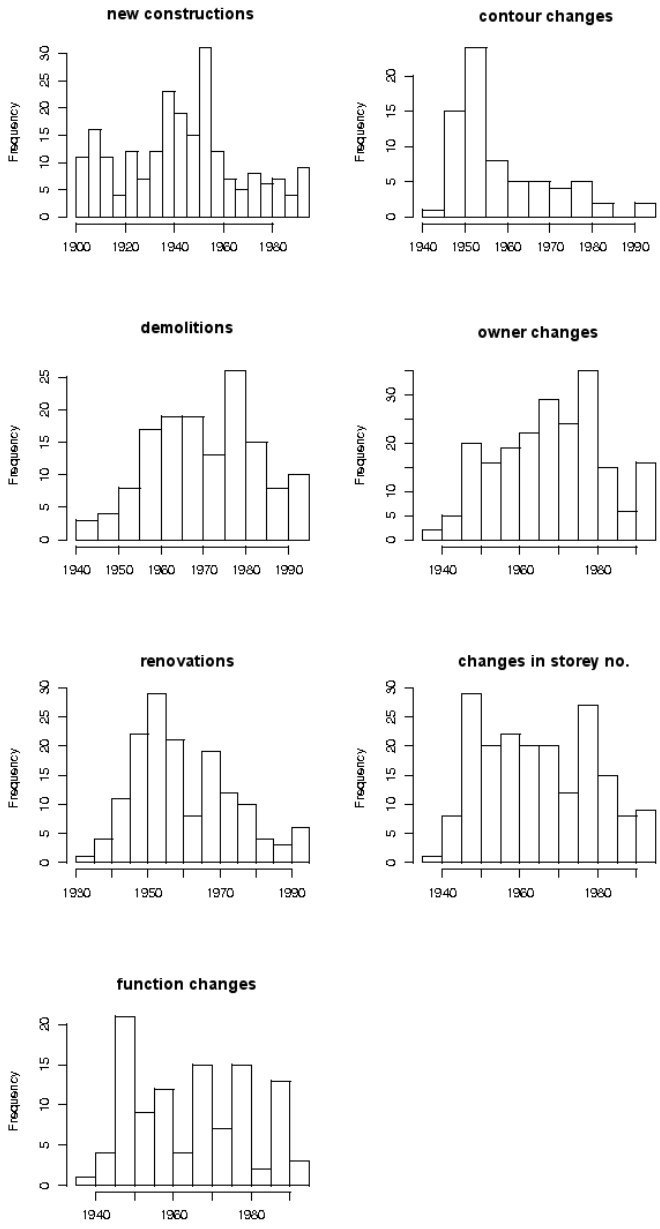

FIGURE 7. Five-year histograms of event counts (event type).

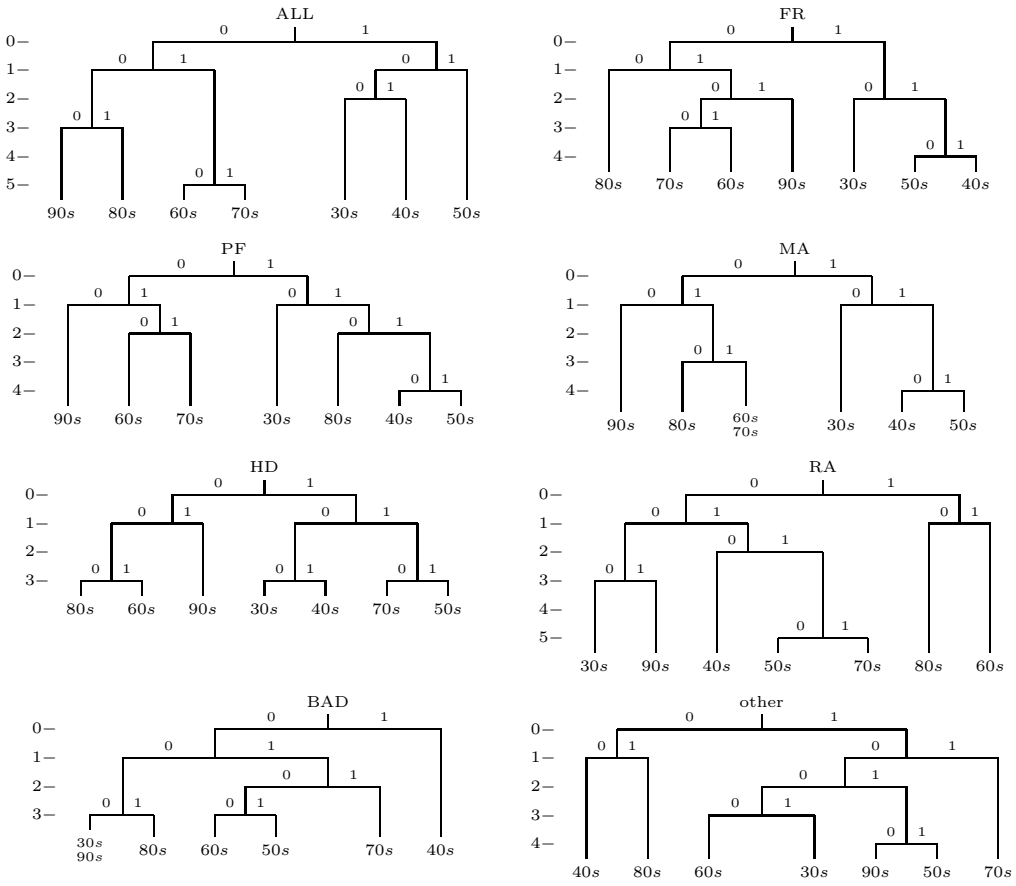

FIGURE 8. Dendrograms of decades by DFT (municipalities).

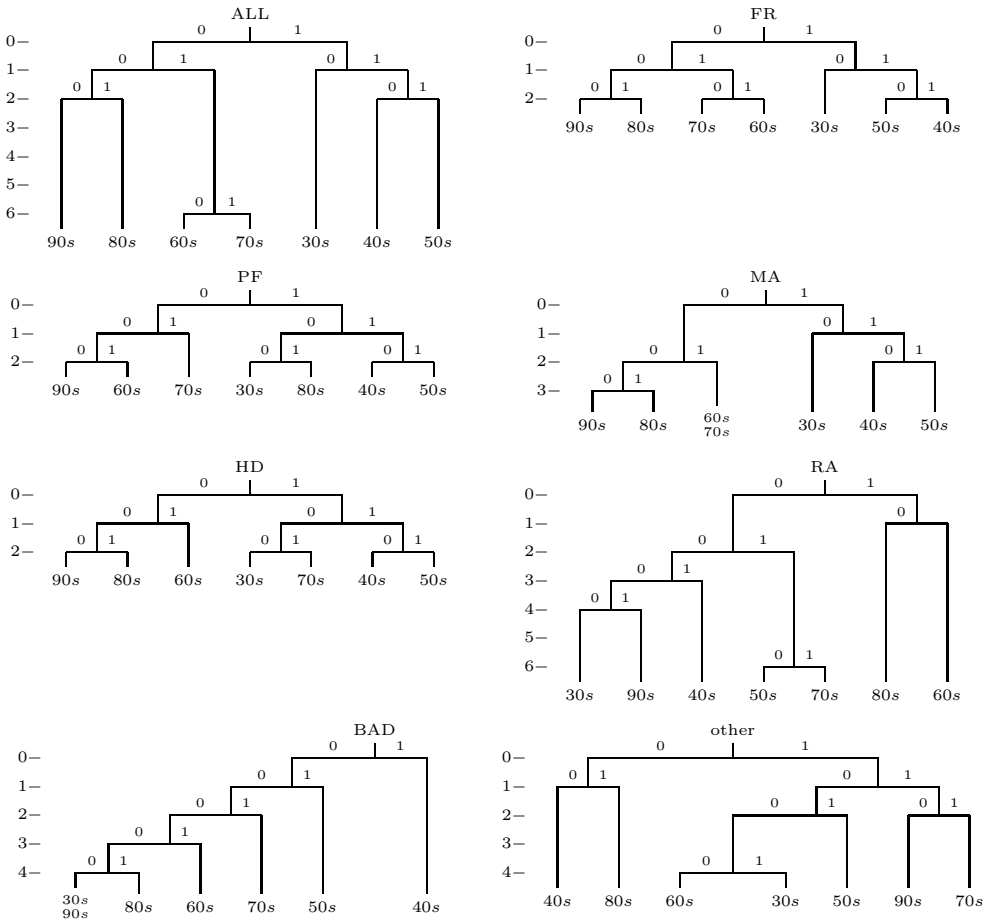

FIGURE 9. Dendrograms of decades by topology (municipalities).

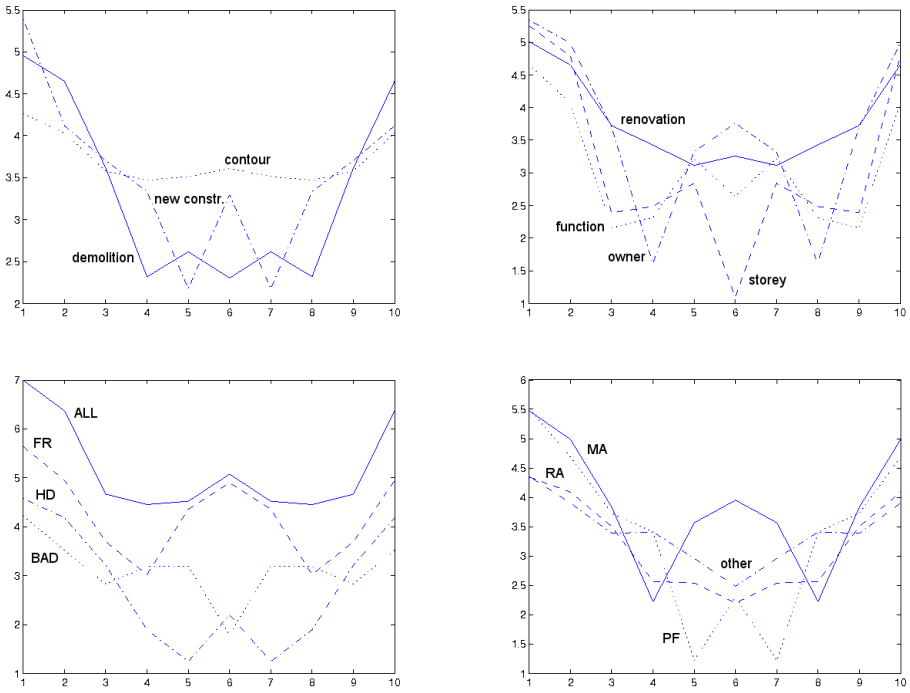

FIGURE 10. Logarithm of absolute of FFT of events (decades). Upper row: count-wise; lower row: municipality-wise.

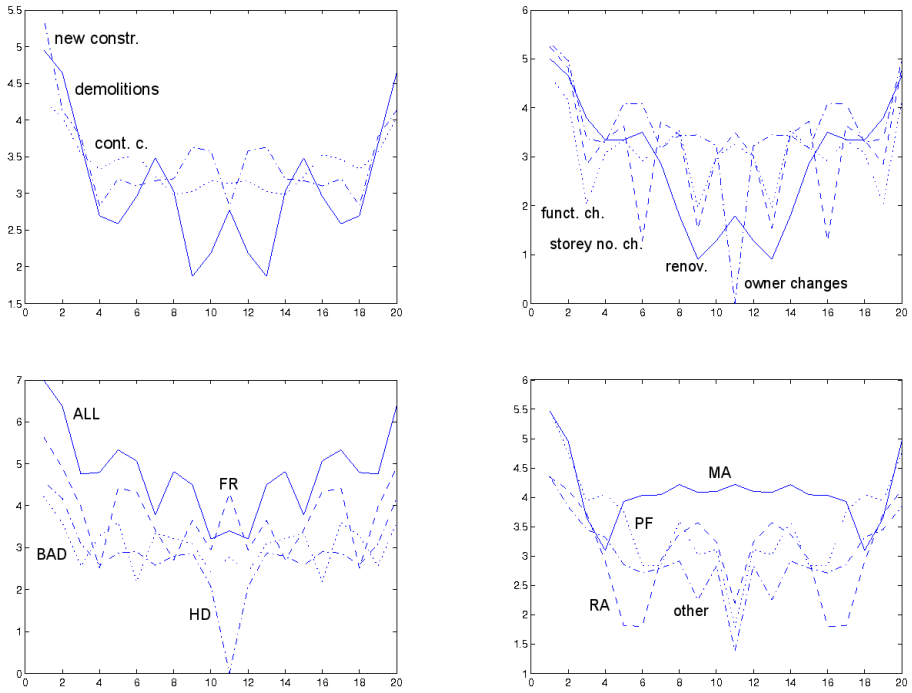

FIGURE 11. Logarithm of absolute of FFT of events (five-year periods). Upper row: count-wise; lower row: municipality-wise.