WOOD 4 sustainability

*p*rocessing

*c*onstruction

*p*roduct

*d*esign

2024

Wood 4 Sustainability

Processing, Construction, Products and Design

2024

Főszerkesztő: Dr. Csiha Csilla

Wood 4 Sustainability

Processing, Construction, Products and Design

2024

SOPRONI EGYETEM KIADÓ

SOPRON, 2024

Közreadó a Soproni Egyetem Faipari Mérnöki és Kreatívipari Kara

Felelős kiadó: Prof. Dr. Fábián Attila

a Soproni Egyetem rektora

Főszerkesztő

Dr. Csiha Csilla Mária

Szerkesztők

Prof. Dr. Kósa Balázs, Prof. Dr. Magoss Endre, Dr. habil. Németh László

Lektorok

Gerencsér Kinga Címzetes Egyetei Tanár, Szabadhegyi Viktor Címzetes Egyetemi Docens,
Prof. Dr. Alpár Tibor, Prof. Emeritus Tolvaj László, Prof. Dr. Markó Balázs

Technikai szerkesztő

Mucsi Zsuzsanna

ISBN 978-963-334-541-2 (pdf)

DOI: https://doi.org/10.35511/978-963-334-541-2

Tartalom

Clustering analysis of the nineteen most populous Hungarian cities using Urban Atlas Building Height Data from Copernicus Land Monitoring Service's 10m raster maps

Zsolt Tóth[a]

[a] University of Sopron, Faculty of Wood Engineering and Creative Industries, 9400 Sopron, Bajcsy-Zsilinszky u. 4. Email: toth.zsolt@uni-sopron.hu

DOI: https://doi.org/10.35511/978-963-334-541-2-18

ABSTRACT

This pilot study delves into the clustering of Hungarian cities based on building height data extracted from Copernicus Building Height 2012 (BH2012) raster maps. BH2012 raster maps have proven overall reliability, yet their potential for city classification and clustering remains underexplored in the literature. While previous studies often prioritize sociological-historical analyses, this research takes a data-driven approach, emphasizing the significance of building height as a fundamental indicator of urban character. By employing four distinct clustering methods, including both legislative criteria and univariate k-means algorithms, this study aims to elucidate underlying patterns in the distribution of building heights across major Hungarian cities. The univariate k-means clustering approach yields more coherent clusters compared to methods based on regulatory frameworks, underscoring the efficacy of data-driven methodologies in capturing nuanced urban characteristics.

Keywords: Clustering, Urban environment, Building height, Data-driven analysis, Hungarian cities, LiDAR raster map, Sustainable cities and communities

1. Introduction

The proportion of buildings of different heights is a significant feature of the character of a city. Copernicus Building Height 2012 (hereafter: BH2012) raster maps of building height may contain minor inaccuracies, but examples from Bratislava (Szatmári et al., 2022) and Warsaw (Kałuski et al., n.d.) show that the raster maps are overall quite accurate and can be used for purposes of urban planning or environmental analyses (Tóth et al., 2024). Sophisticated analyses based on rasterized data layers can be found on the urban climate model of Hong Kong (Kwok et al., 2020), identification of slum areas (Taubenböck & Kraff, 2014), or land assessment (Mansouri Daneshvar et al., 2017) for building height construction among others. However, building heights in raster form can also be used for city classification and clustering. In the clustering of cities, however, this possibility is only tangentially present, as can be seen for example in the otherwise impressive articles on the identification of the types of large cities in Russia (Limonov & Nesena, 2016), dramatic worldwide urbanisation in recent decades (Sun et al., 2020), or spatial typification of the vast urban network of the Yellow River basin (Yin et al., 2023), etc.

2. Theoretical research

There are many examples of the classification of Hungarian cities in the literature. The majority of authors carried out sociological-historical analyses or multidimensional analyses where the selection of criteria is based on mainly a subjective method, which inevitably reflects the social or political conditions of the time of the evaluation (Csomós, 2015).

As in the previous papers, building height usually is only one peripheral factor among many or it plays no role at all (Makra & Sümeghy, 2007). However, this article is based on the premise that the height of buildings and their distribution is a fundamental and informative indicator of a city's character.

The classification of buildings by height is based on height class intervals that vary almost from municipality to municipality, as Hungarian law leaves the building regulations to a large extent to the municipalities. Nevertheless, the National Planning and Building Requirements (in Hungarian abbreviation: OTÉK) gives some guidance on the classification of buildings. OTÉK defines tall buildings (over 30 m high) and medium-tall buildings (between 13.65 and 30 m high). Buildings lower than this are not grouped, so they can be single or multi-storey buildings. Only farmhouses (in Hungarian: "tanyák"), which are common in certain areas of the Hungarian Great Plain, are limited to a maximum height of 4.5 m. But many single-storey buildings in most Hungarian towns are taller than this.

This study is an attempt to create clusters of major cities in Hungary using only rasterized height data of the built environment. In all stages of the analysis process, data science methods are used basically, and the results are embedded in the socio-economic context only in the last stages. The paper can therefore be considered a novelty in its approach and methodology. The assumption is that with the right methodology, a univariate analysis can classify major Hungarian cities surprisingly accurately. An interesting question is to what extent the results obtained confirm or nuance our prior knowledge about Hungarian cities, whose building stock has been significantly influenced by the modernisation of the late 19th century and the rapid industrialisation of the 20th century socialist era, in addition to many medieval (or even older) impacts.

As well as the years after the change of regime (after 1990) have had a minor impact on the building stock of settlements. This varies greatly from one city to another.

3. Material and method

The research based on the BH2012 third version. The dataset is a 10m high resolution raster layer that contains height information for certain cities and metropolitan areas, including data from 19 Hungarian cities with more than 50,000 inhabitants at the time of the survey. Satellite data (LiDAR point clouds) and related datasets, such as the digital surface model (DSM), digital terrain model (DTM), and normalized DSM, are used to generate height information (European Environment Agency, 2022).

From the downloaded 19 Geotiff datasets, RasterLayer objects were defined that can be interpreted in the given environment. The downloaded height data contained rounded values. This did not affect the calculation considerably but gave some caution as to the statistical methods to be used. Values are essentially discrete data that are or can be handled as continuous. However, this category, favoured by statistical software developers, is a source of some controversy. Because of the high number of values, this would be an acceptable method (Clench-Aas et al., 2011), but preferably the values were evaluated as ordinal.

The frequencies of building heights were not really suitable for grouping cities, partly because of rounding to integer and partly because of too many height values. However, appropriate inter-class frequencies may provide an opportunity to do so. The main question is how many class intervals to create and where to draw the boundaries of the class intervals. Four methods have been chosen.

Method 1

The optimal univariate k-means clustering was accomplished in linear time. Using dynamic programming, the employed R function (“*Ckmeans.1d.dp*” from the package with the same name) divides the elements of the numeric vector (in this case of 19 pieces) x into k clusters in the most efficient manner (Wang & Song, 2011) (Song & Zhong, 2020):

$$f(i,k) = \min_{j<i}\big(f(j,k-1) + \mathrm{cost}(j,i)\big), \quad (1)$$

where, $f(i,k)$ = minimum cost for clustering the first i elements into k clusters; $cost(j,i)$ = within-cluster variance for the segment $[j,i]$.

The overall within-cluster sums of squared distances between each element and the associated cluster mean are reduced to a minimum:

$$\mathrm{WCSS} = \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2, \quad (2)$$

where, k = number of clusters; $C_i =$ the i-th cluster; x = a data point in C_i; μ_i = mean of C_i.

The outcome of this algorithm is guaranteed to be optimal and reproducible, in contrast to heuristic k-means algorithms, whose results may not be ideal or may vary from run to run. In that case $k = 3$.

Method 2

Same method as the first method, but $k = 4$.

Method 3

The declared class boundaries of the OTÉK were used $(13.65, 30)$, so three class intervals were defined:
$[minimum\ height = 3, 13.65]$,
$[13.65, 30]$,
$[30, maximum\ height = 199]$.
Because of integer values it is equivalent to $[3, 13]$, $[14, 30]$, $[31, 199]$.

Method 4

Same method as the third method, but we added the $4.5\ m$ building height as a boundary mentioned earlier, so four class intervals were defined:
$[3, 4.5]$, $]4.5, 13.65]$, $]13.65, 30]$, $]30, 199]$.
In that case it is equivalent to $[3, 4]$, $[5, 13]$, $[14, 30]$, $[31, 199]$.

The height values were sorted into class intervals generated by four different methods, and Goodman and Kruskal's gamma values were calculated for each distribution per city pair (Naroll, 1974):

$$\gamma = \frac{P - Q}{P + Q} \quad , \qquad (3)$$

where, P = number of concordant pairs; Q = number of discordant pairs.

This produced four $19x19$ matrices, because we wanted to cluster 19 cities. In any case, for the purposes of our calculations, it is useful to express gamma values in absolute terms. Subsequently, a hierarchical cluster analysis was performed mainly using the R "*hclust*" function based on the values obtained above. The dendrogram's height value of 1 was used to define a soft threshold about or below which real clusters might be possible. The resulting clusters are also displayed on a simplified map. Then, the characteristics of the urban clusters were examined. It was analysed whether the clusters defined by Method 1-2 or the clusters of cities defined by the legislation were more realistic and the possible reasons for the differences were outlined.

4. Results and discussion

The algorithm used in Method 1 have the class intervals set at $[3, 7]$, $[8, 17]$, $[18, 199]$. The class intervals created by algorithm in Method 2 are $[3, 6]$, $[7, 12]$, $[13, 23]$, $[24, 199]$ (Figure 1.) The class intervals in Method 3-4 defined by outer factors and rules as you can see in the previous chapter. The number and distribution of $10x10$ m areas for each cluster (height interval) are shown in Table 1-4 in supplementary materials (Tóth, 2024). The tables mainly show that the drawing of boundaries by different methods could have had a significant impact on the proportions by category. It is questionable what impact this might have on subsequent clusters.

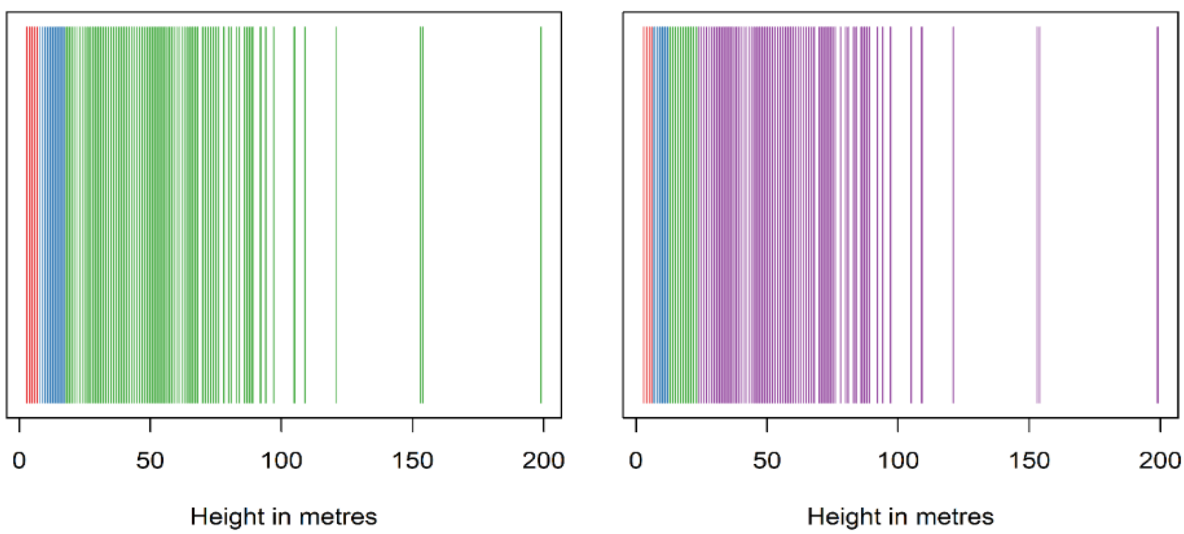

Figure 1. Results of univariate clustering (Method 1-2)

Gamma values (3) in absolute terms can be seen in Table 4-8 in supplementary materials (Tóth, 2024). They are obviously difficult to analyse on their own without a proper clustering procedure. It is clear from the dendrograms that some cities can almost always be grouped into a single cluster, while in other cases the boundaries of clusters are different. It can be clearly seen that the two different clustering methods we used, which differ in the number of height classes created by univariate k-means clustering (1-2), gave essentially identical results. Our results differed from the clusters of cities derived from the legislation, which also showed a divergence from each other (Figure 2-3.).

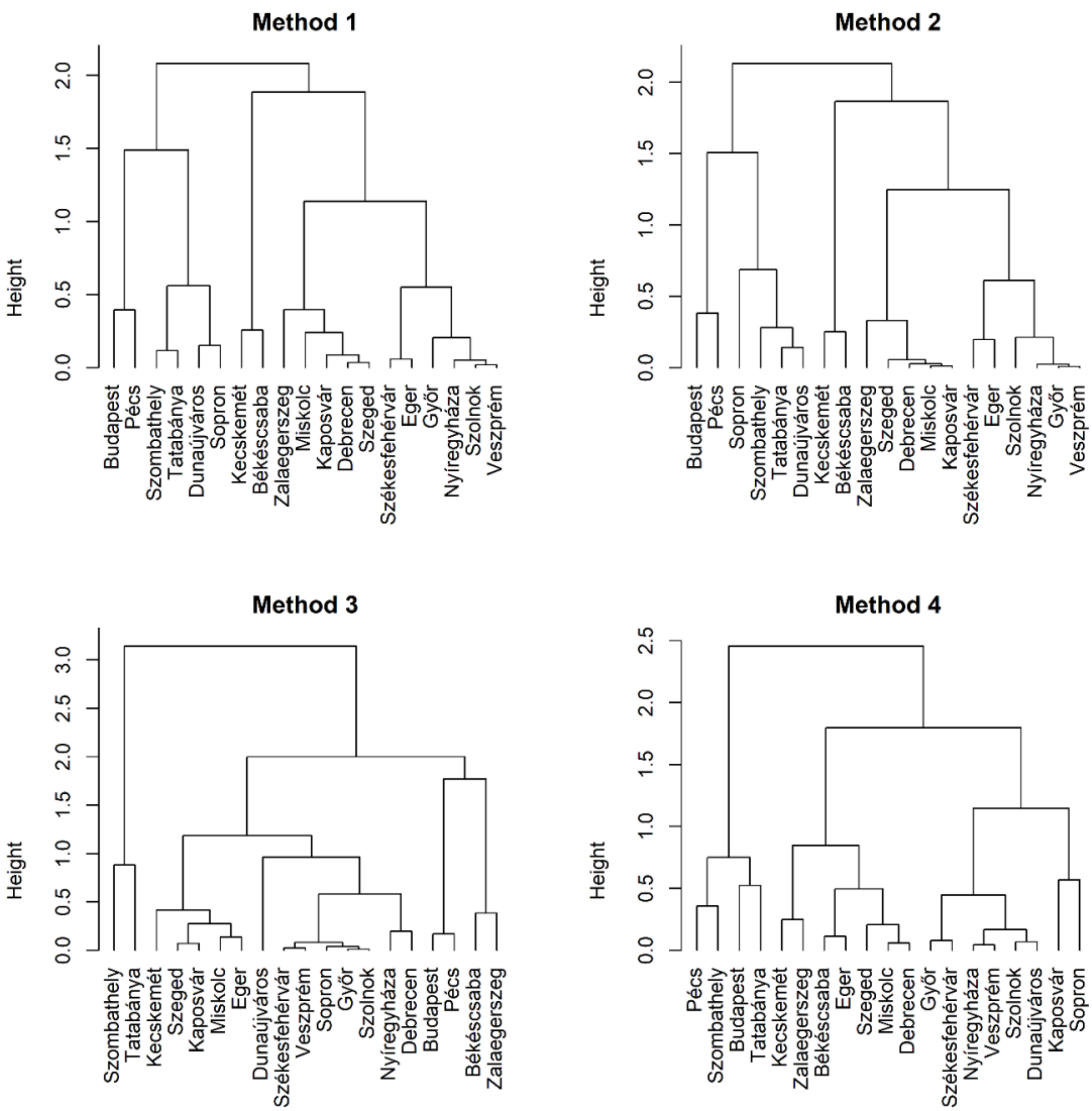

Figure 2. Dendrograms of four methods

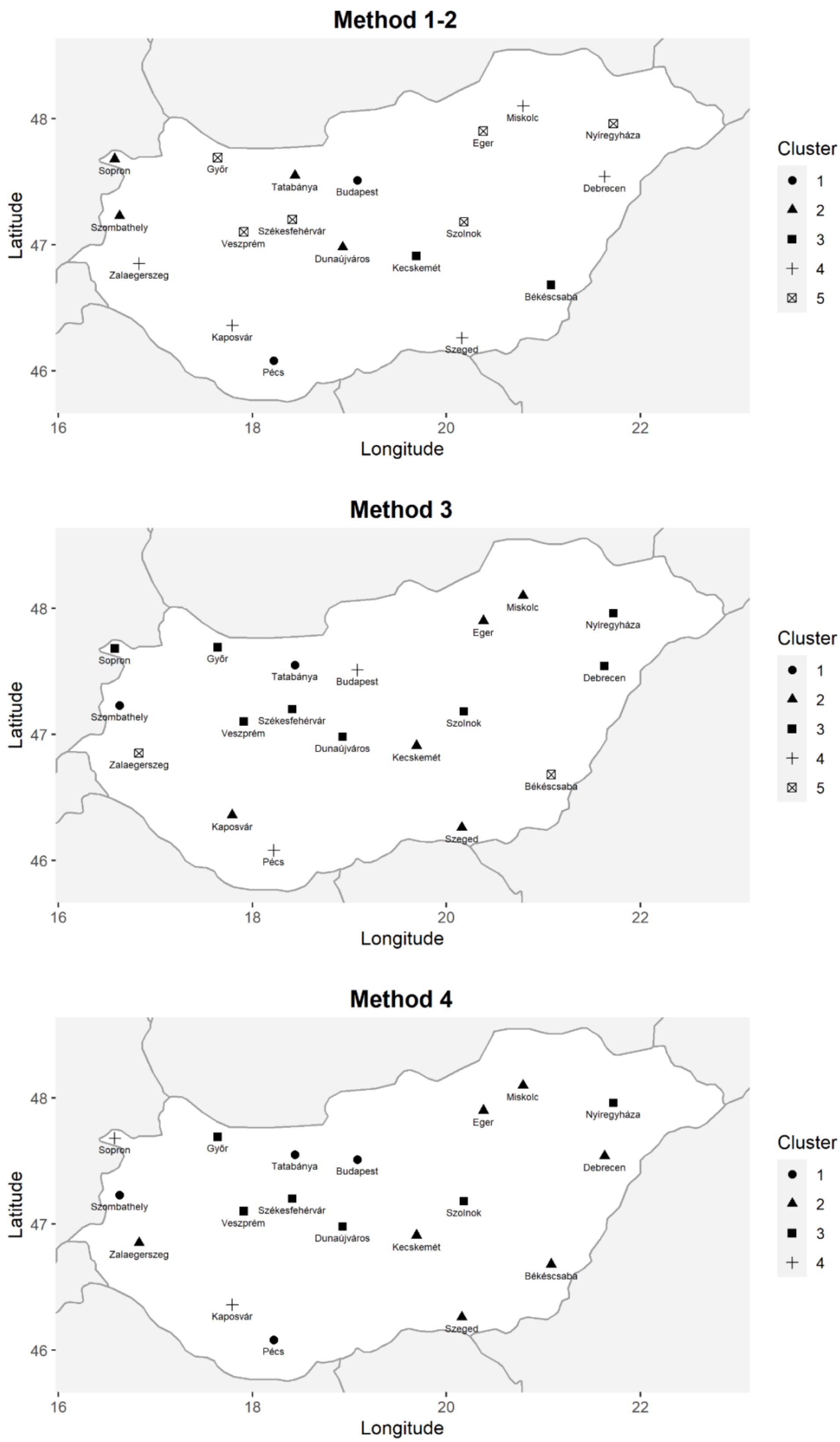

Figure 3. Clusters on map

There are probably several reasons for this discrepancy, but arguably the main reason is that the ten-storey houses ("block of flats" or "panel buildings") built using the factory method during the socialist era were placed in different height categories according to our methods and the methods based on the OTÉK.

Ten-storey buildings are usually around $30-35$ metres in height, depending on the construction technique, and as sometimes only part of them fell within the $10x10$ metre area, many prefabricated housing areas fell below 30 metres in the survey.

The classification by methods 1-2 probably put most of the prefabricated houses into one class, but methods 3-4 did not, with the upper class boundary bisecting the range of 10-storey houses. Similar anomalies can be found for the other class midpoints, as the clusters of cities created by methods 1-2 placed cities with very different histories, but sometimes with surprising similarities, in the same cluster.

A good example of this is the case of Sopron and Szombathely, which can be considered as historical cities, and Tatabánya and Dunaújváros, which are industrial cities in the same cluster. In the first half of the socialist era, the centres of Tatabánya and Dunaújváros were characterised by the construction of a number of four-storey Stalinist-baroque buildings, which, along with a variety of other architectural styles, created a similar inner city to the historic towns, alongside the existing suburbs.

And during socialism, all four cities saw the construction of a large number of four and ten-storey houses, which created similar urban structures in very different traditions.

5. Conclusions

The analysis of cities within the clusters, and specifically the clusters themselves, extends beyond the scope of this article. However, it is evident that Method 1-2 has demonstrated superior effectiveness as a clustering procedure compared to the approach based on regulatory-legal environments, which purportedly integrate historical and contemporary experiences. This suggests that the regulatory framework may not necessarily be antiquated; rather, it is conceivable that the univariate k-means clustering method is better suited to address the discrepancy arising from the imprecise delineation of individual building outlines within the 10x10 meter units.

A more precise clustering outcome could potentially be attained by incorporating population density data from all urban settlements, particularly if augmented with highly accurate LiDAR (ALS) data and subjected to advanced data science methodologies. However, the current availability of such comprehensive data sources and analytical techniques remains limited in Hungary.

Future directions for research in this domain may involve exploring alternative clustering algorithms that can better accommodate the complex spatial dynamics inherent in urban environments. Additionally, efforts to enhance data collection infrastructure, such as expanding access to LiDAR technology and promoting collaboration among research institutions, could facilitate more comprehensive analyses and foster advancements in urban clustering methodologies. Furthermore, integrating socio-economic indicators alongside geographical data may offer deeper insights into the underlying patterns of urban development and population distribution, thereby enriching the precision and applicability of clustering analyses in urban studies.

The analysis based on univariate k-means algorithms obviously has the limitations of univariate analyses. In this case, it worked better than the height classification resulting from the legal regulation, but there were technical reasons for this. The method is not generalisable to all cases, clustering based on in-depth sociological research may be better than the method used. But the clustering procedure performed quite well here.

Author statements

Zsolt Tóth was the sole author and carried out all activities.

Data availability

This publication has been prepared using European Union's Copernicus Land Monitoring Service information: https://doi.org/10.2909/42690e05-edf4-43fc-8020-33e130f62023.

References

Clench-Aas, J., Nes, R.B., Dalgard, O.S. & Aarø, L.E. 2011. Dimensionality and measurement invariance in the Satisfaction with Life Scale in Norway. *Quality of Life Research*, 20(8): 1307–1317. https://doi.org/10.1007/s11136-011-9859-x

Csomós, G. 2015. The ranking of cities as centres of the Hungarian economy, 1992–2012. *Regional Statistics*, (01): 66–85. https://doi.org/10.15196/RS05104

European Environment Agency. 2022. *Urban Atlas Building Height 2012* (raster 10 m), Europe - version 3, Oct. 2022 (03.00). https://doi.org/10.2909/42690e05-edf4-43fc-8020-33e130f62023

Kałuski, M., Hościło, A. & Gurdak, R., 2012. *Accuracy assessment of the Copernicus Buildings Height 2012 layer based on the city of Warsaw*. http://bc.igik.edu.pl/dlibra/docmetadata?id=872

Kwok, Y.T., De Munck, C., Schoetter, R., Ren, C. & Lau, K.K.-L., 2020. Refined dataset to describe the complex urban environment of Hong Kong for urban climate modelling studies at the mesoscale. *Theoretical and Applied Climatology*, 142(1): 129–150. https://doi.org/10.1007/s00704-020-03298-x

Limonov, L.E. & Nesena, M.V., 2016. Structural and economic typology of large Russian cities. *Regional Research of Russia*, 6(2): 144–155. https://doi.org/10.1134/S2079970516020052

Makra, L. & Sümeghy, Z., 2007. Objective analysis and ranking of Hungarian cities, with different classification techniques, Part 2: Analysis. *Acta Climatologica et Chorologica Univ. Szegediensis*, (40–41): 91–100.

Mansouri Daneshvar, M.R., Khatami, F. & Shirvani, S., 2017. GIS-based land suitability evaluation for building height construction using an analytical process in the Mashhad city, NE Iran. *Modeling Earth Systems and Environment*, 3(1): 16.
https://doi.org/10.1007/s40808-017-0286-z

Naroll, R., 1974. An Exact Test of Significance for Goodman and Kruslkal's Gamma. *Behavior Science Research*, 9(1): 27–40.
https://doi.org/10.1177/106939717400900112

Song, M. & Zhong, H., 2020. Efficient weighted univariate clustering maps outstanding dysregulated genomic zones in human cancers. *Bioinformatics*, 36(20): 5027–5036.
https://doi.org/10.1093/bioinformatics/btaa613

Sun, L., Chen, J., Li, Q. & Huang, D., 2020. Dramatic uneven urbanization of large cities throughout the world in recent decades. *Nature Communications*, 11(1): 5366.
https://doi.org/10.1038/s41467-020-19158-1

Szatmári, D., Kopecká, M. & Feranec, J., 2022. Accuracy Assessment of the Building Height Copernicus Data Layer: A Case Study of Bratislava, Slovakia. *Land*, 11(4): 590.
https://doi.org/10.3390/land11040590

Taubenböck, H. & Kraff, N.J., 2014. The physical face of slums: a structural comparison of slums in Mumbai, India, based on remotely sensed data. *Journal of Housing and the Built Environment*, 29(1): 15–38.
https://doi.org/10.1007/s10901-013-9333-x

Tóth, Z., 2024. *Supplementary materials for paper* 'Clustering analysis of the nineteen most populous Hungarian cities using Urban Atlas Building Height Data from Copernicus Land Monitoring Service's 10m raster maps'.
https://zenodo.org/doi/10.5281/zenodo.14278443

Tóth Z., Farkas P. & Novotni A., 2024. Individual tree detection and spatial distribution analysis without reference data. *Dimenziók*, 12(12): 33–41.
https://doi.org/10.20312/dim.2024.04

Wang, H. & Song, M., 2011. Ckmeans.1d.dp: Optimal k-means Clustering in One Dimension by Dynamic Programming. *The R Journal*, 3(2): 29–33.
https://doi.org/10.32614/RJ-2011-015

Yin, Y., Peng, W. & Xie, X., 2023. Spatial–temporal characteristics and obstacle factors analysis of urban resources and environment carrying capacity in the Yellow River Basin cities. *Environment, Development and Sustainability*.
https://doi.org/10.1007/s10668-023-04040-z