ACTA CLIMATOLOGICA ET CHOROLOGICA
Universitatis Szegediensis, Tom. 38-39, 2005, 59-69.

MODELLING THE MAXIMUM DEVELOPMENT OF URBAN HEAT ISLAND WITH THE APPLICATION OF GIS BASED SURFACE PARAMETERS IN SZEGED (PART 2): STRATIFIED SAMPLING AND THE STATISTICAL MODEL

T. GÁL[1], B. BALÁZS[1] and J. GEIGER[2]

[1] *Department of Climatology and Landscape Ecology, University of Szeged, P.O.Box 653, 6701 Szeged, Hungary*
E-mail: tgal@geo.u-szeged.hu
[2]*Department of Geology and Paleontology, University of Szeged, P.O.Box 658, 6701 Szeged, Hungary*

Összefoglalás – Az urbanizált környezet lokális léptékű klímamódosulást eredményez a városok területén, amelynek legszembetűnőbb megnyilvánulása a magasabb hőmérséklet, az ún. városi hősziget. Kutatásunk célja az, hogy a városi felszínparaméterek, illetve matematikai-statisztikai módszerek segítségével az éves átlagos maximális hősziget intenzitására becslést készítsünk. A számításokban olyan új paramétereket is felhasználtunk, amelyek a város geometriáját három dimenzióban jellemzik. A felmérési módszer munkaigénye miatt a várost szerkezetében és térben reprezentáló, a teljes terület egyharmadára kiterjedő mintaterületen végeztük el vizsgálatunkat. Kiválasztását az ún. rétegzett mintavételezési eljárással készítettük. A statisztikai vizsgálataink azt bizonyítják, hogy a kompaktsági mutató a széles körben elfogadott égboltláthatóság paraméternél is erősebb kapcsolatot mutat a hősziget intenzitással. Lépésenkénti lineáris regressziós eljárás segítségével alkottuk meg azt a modellt, amellyel – felhasználva a területről származó különféle felszínparamétereket – már becslést készíthetünk a városi hősziget területi szerkezetére. Az eredmények azt mutatják, hogy a modellezett hőszigeti mező képe csak kis mértékben tér el a valóságos állapottól, ami bizonyítja az új paraméterek jelentőségét, és a mintaterület kiválasztásának helyességét.

Summary – Our investigations concentrated on the urban heat island (*UHI*) in its strongest development during the diurnal cycle in Szeged, Hungary. In order to quantify the effect of the peculiar urban structure on the development of the mean annual urban heat island we determined a new surface parameter (weighted volumetric compactness) which characterises the volume, the building plan area and the thermodynamic role of the buildings at the same time. The calculation of this new parameter required a large-sized digital database that includes more than 22,000 building's 3 dimensional measurement. Because this would take a long time, we concentrate the investigation on a smaller but representative sample area, as the first step of our research. Our task included the development of statistical models using urban surface parameters (built-up and water surface ratios, sky view factor, building height, weighted volumetric compactness). Model equations were determined by means of stepwise multiple linear regression analysis. As the results show, there is a clear connection between the spatial distribution of the *UHI* and the examined parameters (built-up and water surface ratios and weighted volumetric compactness), so these parameters play an important role in the evolution of the *UHI* intensity field. The distribution of the difference between the modelled and the (independent) annual mean maximum heat island intensity show that we could calculate the heat island's spatial distribution properly from the sample area's dataset.

Key words: Urban heat island, urban surface parameters, weighted volumetric compactness, representative sample area, stratified sampling, stepwise multiple linear regression model, Szeged, Hungary

1. INTRODUCTION

The first part of this paper (*Balázs et al.,* 2005) and the earlier studies have described the investigation area, and the method of measuring the temperature and surface parameters

(*Sümeghy and Unger,* 2003; *Unger,* 2004). In order to quantify the effect of the peculiar urban structure on the development of the mean urban heat island a new surface parameter (weighted volumetric compactness) was determined which characterises the volume, structure and thermodynamic role of buildings at the same time. In this paper, we use this dataset, but disregard the detailed description of the measuring methods. In order to study microclimate alterations within the city, the utilization of statistical modelling may provide useful quantitative information about the spatial and temporal features of the urban temperature excess by employing different surface parameters (*Oke*, 1981).

Our purpose is to investigate the quantitative effects of the relevant surface parameters on the *UHI* patterns. These factors are: the built-up ratio, the water surface ratio and the sky view factor. Our task is to prove that the connection between the compactness and the annual *UHI* intensity is significant, and we would also demonstrate that this new parameter is a useful part of our statistical model.

2. STUDY AREA AND METHODS

2.1. Selection of the sample area

As we already mentioned, in our project we plan the measuring of the characteristic geometrical and morphological parameters in the whole area. It is important because we would like to determine the connection between these parameters and the *UHI* intensity. Such a detailed and large-scale analysis of urban geometry – as far as we know – is without precedent in the region.

In the investigated area the number of the houses is more than 22,000. Presumably, it would have taken too much time to measure the parameters of this enormous number of buildings. Therefore, we decided on conducting our research in a smaller area. To prove that these parameters have a significant role in developing of the *UHI*, we did our research in a representative sample area including 35 cells from the 107-cell grid network. This makes statistical investigation possible and less time is needed for the measurements.

2.1.1. Stratified sampling

Stratified sampling is a sampling design in which prior information about the population is used to determine groups (called strata) that are sampled independently. Each possible sampling unit or population member belongs to exactly one stratum. There can be no sampling units that do not belong to any of the strata and sampling units that belong to more than one stratum. When the strata are constructed to be relatively homogeneous with respect to the variable being estimated, a stratified sampling design can produce estimates of overall population parameters (e. g. mean, proportion) with greater precision than estimates obtained from simple random sampling.

If the investigator has prior knowledge of the spatial distribution of the study area, the strata should be defined so that the area within each stratum is as homogeneous as possible. The variable providing the information used to establish the strata (the so-called „auxiliary variable") was the built-up ratio.

The fact, that the increase in precision depends on the strength of the correlation between the auxiliary variable and the outcome variable to be estimated, may theoretically be a restricting factor. If there is not any significant correlation between the auxiliary

variable and the one being estimated, the precision of the final estimation can be significantly decreased.

The strata should be determined before allocating the sample sizes. When the strata are defined according to an auxiliary variable that is correlated with the variable to be estimated, the optimal definition of the strata is that the population included in each stratum should be as homogeneous as possible with respect to the auxiliary variable.

Cochran (1963) offers some guidelines on how to optimally assign strata when the auxiliary variable is continuous. If there is a particular interest of estimating the overall mean for the population, Cochran suggests defining no more than six strata and using a procedure attributed to *Dalenius and Hodges* (1959) to determine the optimal cutoff values for each of the strata based on the distribution of the auxiliary variable for the population. In this study six layers have been defined and the samples have been arranged into layers by applying the method of *Dalenius and Hodges* (1959).

Table 1 shows the final result of defining the six layers.

Table 1 Number of cells in each layer, and the number of cells by layers

Strata	Cutoff values (built up ratio in %)	In layers	In samples
#1	$B \leq 25.84$	11	4
#2	$25.84 < B \leq 43.56$	14	5
#3	$43.56 < B \leq 58.80$	18	6
#4	$58.80 < B \leq 71.68$	20	7
#5	$71.68 < B \leq 83.60$	22	7
#6	$83.60 < B$	18	6
Total		*103*	*35*

2.1.2. Selection by the spatial distribution

With the stratified sampling we allocated the number of the samples from each strata. After that the random selection of the cells could be an adequate solution, but if we have considered other aspects the outcome would be better. We had two considerations. First of all we decided that the cells should be distributed evenly. Furthermore, we had to choose the cells where the horizontal thermal gradient was high when the *UHI* developed.

From the possible variations we have chosen the one whereof we could interpolate the spatial distribution of the *UHI* with the minimum deviation. As a result of this process, cells of the chosen sample area are relatively well scattered all over the research area (*Fig. 1*). In those places where the chosen cells are located near each other, the horizontal temperature gradient is high in the time of the development of the *UHI*.

2.2. Construction of the stepwise multiple linear regression model

In order to assess the extent of the relationships between the annual mean maximum *UHI* intensity (ΔT) and various urban surface factors, multiple linear regression analyses were applied. To determine model equations we used ΔT as predictant (dependent variable) and the afore mentioned parameters as predictors: ratios of built-up surface (B) water surface (W), mean sky view factor (SVF), average compactness (C_m) and weighted volumetric compactness (C_v) by cells.

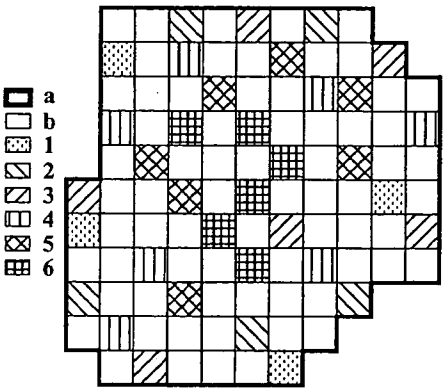

Fig. 1 Distribution of the cells of the sample area in the investigated area: (a) border of the investigated area, (b) cells not included in the sample area, (1-6) cells selected from the given layers

According to the results of previous studies the connection between the urban surface parameters (*B*, *W*, *SVF* and *H* = building height) and the mean *UHI* intensity is well-describable by a linear function (*Bottyán and Unger,* 2003; *Bottyán et al.,* 2003). Thus, constructing the present model, we also applied the linear-based approach.

We presumed that all five parameters have significant impact on the spatial distribution of the *UHI*. First, all factors were included in the database (as predictors). We selected the statistically acceptable predictors, later applied in the model, by the stepwise linear regression method. In the process we applied the SPSS for Windows 9 software. Limits of predictors were entered or removed from the model depending on the significance of the F value of 0.01 and 0.05, respectively.

3. RESULTS AND DISCUSSION

3.1. The representativity of the sample area

Because of the stratified sampling being based on the built-up ratio, we examine the sampling method's errors with this parameter. Because of the sampling method we only have to examine how representative the sampling is for the spatial distribution of the whole dataset.

It would not be appropriate to study the spatial representativity of the selection by using the spatial distribution of the interpolated values of the built-up ratio, since none of the interpolation methods are suitable for the spatial extension of such a rhapsodically changing parameter. Earlier studies proved the significant connection between the built-up ratio and the *UHI* intensity (e.g. *Unger et al.,* 2000). Therefore we examine the representativity by the application of the *UHI* field. We interpolated the spatial distribution of the *UHI* intensity based on the complete database, and also based on the data of the selected 35 cells. The later version of the heat island field is more simplified and the run of isotherms is more settled, less detailed, but in its main characteristics it is basically similar to the *UHI* field based on the complete database (*Fig. 2a-b*).

Neglecting the smaller structural characteristics, only small-scale differences occur, and in case of three-fourth of the area this difference does not even reach 0.1°C (*Fig. 3*). The mean of the differences is -0.035°C, the standard deviation is 0.11°C. In order to identify the value significantly different from the mean deviation, the limits of confidence interval belonging to the database had to be calculated (at 5% significance level). Thus, we can find those areas where the error is especially great, namely the cases where the selection was not appropriate. Significant positive differences occupy 0.1% of the area, negative differences occupy 3% of the area.

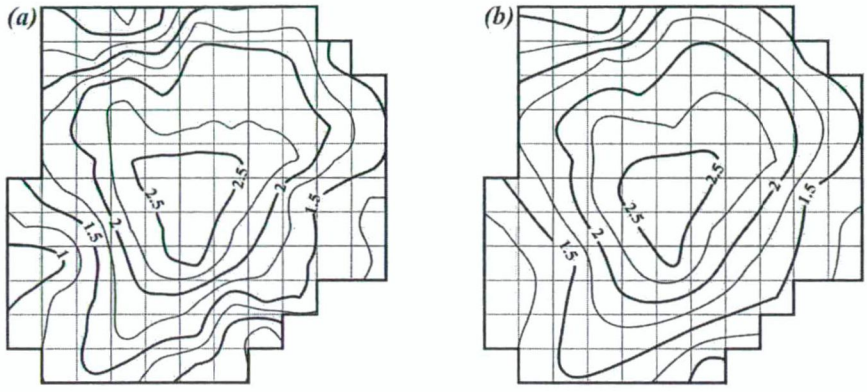

Fig. 2 Spatial distribution of the annual mean maximum urban heat island intensity (a) based on the complete database and (b) based on the data of the selected 35 cells (April 2002 – March 2003)

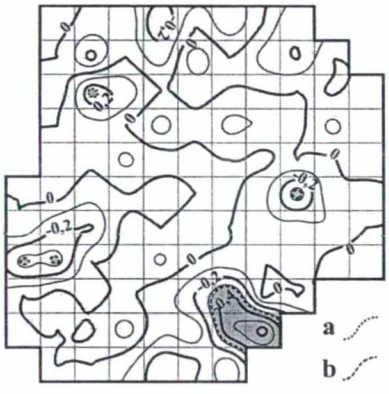

Fig. 3 The distribution of the difference between the annual mean maximum urban heat island intensity based on the complete database and based on the data of the selected 35 cells (April 2002 – March 2003), a - the upper limit of the confidence interval (0.25°C), (b) the lower limit of the confidence interval (-0.32°C)

The small values of deviation prove that the two structures are mainly similar and therefore the selection process was successful. Nevertheless, it is still important to take into account the afore-mentioned deviations because they present the maximum accuracy of the heat island field based on the later-discussed model equation.

3.2. The relationships between the surface parameters and the UHI

Firstly we started to assess the relationships between the mean maximum *UHI* and the surface parameters, we analysed the connection pairs-wise using the data from 35 cells. During this process we identified the formulae (according to the $y = ax + b$ general form) of linear regression lines referring to the closeness of the stochastic connection between each parameter and the ΔT as well as the values of the determination coefficient (r^2) and the values of standard deviation around the regression line (σ_R).

The null-hypothesis, namely that there is no real connection between two chosen parameters, can arise only in those cases where the value of the determination coefficient is large enough. The acceptancy interval of the null-hypothesis, in case of 35 elements on 5% significance level was $r^2 > 0.1089$ (*Péczely*, 1979).

Table 2 The statistical relationships between the *UHI* intensity and the surface parameters

Parameter	Linear regression equation	r^2	σ_R
B	$\Delta T = 0.0165 \cdot B + 0.8005$	0.6619	0.293°C
W	$\Delta T = -0.0019 \cdot W + 1.7965$	0.0011	0.5041°C
SVF	$\Delta T = -3.4219 \cdot SVF + 4.8782$	0.3584	0.404°C
C_m	$\Delta T = 0.363 \cdot C_m + 0.9351$	0.2976	0.423°C
C_v	$\Delta T = 2 \cdot 10^{-7} \cdot C_v + 1.4827$	0.5243	0.348°C

Based on the statistical parameters (r^2, σ_R) we can pronounce that the closest connection is between the values of the B and the ΔT. The trend is positive, so when the value of the built-up ratio is increasing the *UHI* intensity is increasing too. (*Table 2*) This is not a surprising result as the main reason of choosing this sample area was exactly the above-mentioned parameter.

The trend is negative between the *SVF and* ΔT, so when the value of the sky view factor is increasing the *UHI* intensity is decreasing. The connection between the *SVF* and the ΔT is statistically significant (*Table 2*), although the value of the r^2 was a bit under our expectations based on previous research (*e.g. Oke* 1981; *Unger* 2004).

There is not close connection between the C_m and the ΔT, however there is real connection based on the determination coefficient. The trend is positive so if the value of C_m is increasing the ΔT will increase at the same time (*Table 2*).

Our hypothesis, namely that C_v parameter is an essential factor in the heat-island development and therefore there is strong stochastic connection between them, was proved by the previously-done correlation examination (*Table 2*). Based on the regression equation, it can be stated that with the increase of the C_v values the temperature difference undoubtedly grows, too (*Fig. 4*). The recently introduced determination coefficient belonging to the C_v parameter is close to the value of the built-up ratio. On the basis of these preliminary results we can conclude that in the explanation of the mean maximum *UHI* intensity structure the C_v parameter carries significantly more information than the widely-accepted surface-parameter, the *SVF*.

3.3. The results of stepwise multiple linear regression

By the application of the above-mentioned method, out of the five original predictors three were statistically acceptable for the estimation of the *UHI* intensity (*Table 3*). The importance of these three parameters in the development of temperature excess was almost 80% ($r^2 = 0.786$). The model is acceptable even on a significance level less than 0.1%, and thus the estimation based on this model is highly reliable. This fact clearly shows that by entering the afore-mentioned parameters, the increase of the value of the explained determination coefficient (r^2) decreases stepwise. The entering of the C_v parameter resulted in a 9.2% increase in the explained correlation, then predictor W adds to this value a further

3.2%. The application of the fourth and fifth parameters (*SVF*, C_m) does not provide more information to the model in practice, and thus, they can be discluded from the model. In case of the *SVF* this fact is quite surpising, because – according to some earlier studies (*Bottyán and Unger*, 2003; *Unger et al.*, 2004) – strong correlation was detected between the *SVF* and the ΔT. This can be explained by the probable fact that it is in multi-collinearity with the C_v, as both parameters are referring to the vertical structure of the town, and therefore only the stronger predictor appears in the model. The fact that the C_v, is a strong predictor in the model confirms our previous theory based on physical experience.

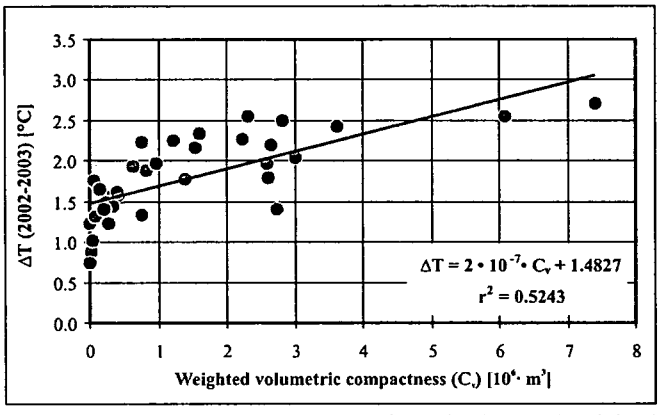

Fig. 4 Relationship between the annual mean *UHI* intensity (ΔT) and weighted volumetric compactness (C_v) (n = 35)

Table 3 Values of the stepwise correlation of mean maximum *UHI* intensity and urban surface parameters, as well as their significance levels in Szeged (n = 35)

Parameter entered	Multiple |r|	Multiple r^2	Δr^2	Significance level
B	0.814	0.662	0.000	< 0.001
B, C_v	0.869	0.754	0.092	< 0.001
B, C_v, *W*	0.886	0.786	0.031	< 0.001

Table 4 Values of coefficients, standard errors and significance levels of the applied urban surface parameters of the models in Szeged (n = 35)

Parameter	Coefficients	Standard deviations	Significance level
B	$1.332 \cdot 10^{-2}$	0.002	< 0.001
C_v	$1.045 \cdot 10^{-7}$	0.000	0.002
W	$1.082 \cdot 10^{-2}$	0.005	0.041
Constant	0.809	0.123	< 0.001

Afterwards, on the basis of the sample data, an estimation is given for the value of the regression model coefficients (*Table 4*). This is important because in case of known coefficients the model-equation can be described. By this equation it is possible to estimate the heat island intensity of the cells and thus spatial structures can be constructed. It appears in the *Table 4* that the estimation of coefficients is especially good, as significancy values

are above 95% in all case. What is more, this value is smaller than 99% only in case of W parameter. The model-equation is calculated as follows:

$$\Delta T = 1.332 \cdot 10^{-2} \cdot B + 1.045 \cdot 10^{-7} \cdot C_v + 1.082 \cdot 10^{-2} \cdot W + 0.809$$

With the application of the equation, it is possible to provide the estimated value of any of the 35 cells. In this statistical model, special attention must be paid to the problem of extensibility, namely that the model can be applied only to parameters with values between the minimum and maximum values applied in its creation. In this case, however, the above-mentioned fact does not have any determining effects. Nevertheless, when applying the model to another town, it has to be considered whether the used predictors are within the adequate intervals. As the limits of extensibility, intervals are given in *Table 5*.

Table 5 The maximal and minimal values of predictors in the model

Parameter	minimal values	maximal values
B (%)	3.24	93.8
C_v (m^3)	1849.54	7411700.05
W (%)	0	40.36

With the help of the Kriging interpolation method (linear variogram-model application), the already-calculated ΔT values provided a basis to the spatial extension of the above-mentioned values. Using this extension, it is possible to define the spatial structure of the *UHI* intensity and thus the whole mean heat island can be detected, practically without any temperature measurements (*Fig. 5a*). Naturally, it is useful to test the model and thus, to compare the ΔT field calculated by the model-equation to an independent database collected in another time period.

3.4. The results of spatial extension and model-verification

We studied the accuracy of the heat island field estimated by the model-equation (*Fig. 5a*) in a number of steps. The independent temperature measurements, which took place between March 1999 and February 2000, were taken as reference data referring to the calculated heat island intensity values of the town.

The first step was to calculate the difference between intensity values of the heat island estimated by the model-equation and real intensity data interpolated from the values of the sample area (*Fig. 5b*). On the basis of these differences we can conclude that the model overestimates real values: the mean deviation is 0.22°C. The absolute deviation, which is smaller than 0.1°C, extends to more than one-third of the whole study area.

With the help of the determined bounds of the confidence interval belonging to the data set we can recognise the areas with especially striking errors. In these areas the estimation of the model significantly differs from the value of mean error. In the estimation of the model, greater negative errors appeared at less than 1% of the study area, while in case of positive ones, this area-ratio was around 3%.

The difference map (*Fig. 5b*) shows those places where the result of modelling was not entirely acceptable. These deviations can be explained by the fact that the model does not take into account values of neighbouring cells that is the equalising effect, which indirectly appears in the given cell while creating the measured *UHI* intensity. In the area of

greater positive deviations the problem is that the estimation gives the same values to the cells of suburbs than to the densely built-up downtown cells. There is only one cell with negative deviation: this is presumably caused by the effect of the river Tisza and thus the values of surface parameters decrease. Since temperature is not able to follow this sudden change real temperature must be higher than the estimated one.

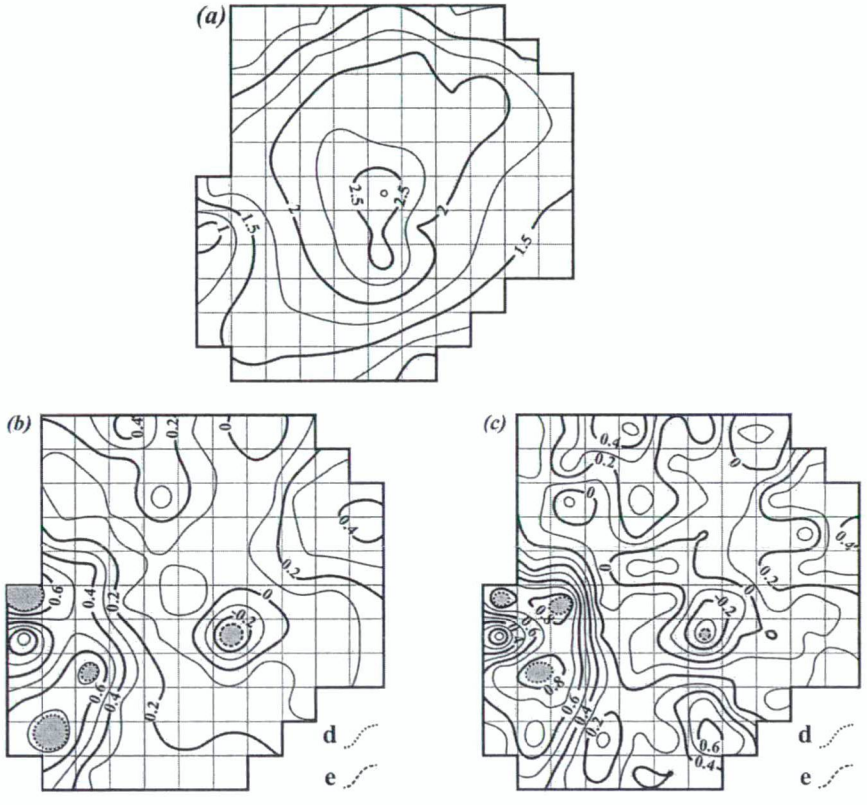

Fig. 5 (a) The modelled heat island distribution, the distributions of the difference between the modelled and the (independent) annual mean maximum heat island intensity (March 1999 – February 2000) (°C) (b) based on the data of the selected 35 cells and (c) based on the complete study area, lower limit of confidence interval is -0.36°C, higher limit is 0.84°C. The boundary of significant (d) negative and (e) positive errors

Afterwards we calculated differences between the heat island intensity values estimated by the model-equation and the real heat island intensity values of the entire study area of 103 cells (*Fig. 5c*). On the basis of these differences, the mean error of the model and the extension was 0.24°C, which is not much more than the error of the model in itself. The statistically large-sized error extended only to a part of the area, smaller than in case of the above-mentioned difference. Errors appearing in the previous case are the results of the selection, although they are much less notable in size.

Taking all these factors into account, we can state that values estimated by the model are closely related to the independent values of the above-mentioned series. In the analysis

of deviations, we have to consider the fact that on the basis of data taken from the sample area a pattern of the whole heat island can be provided with few errors. The spatial structure of larger differences points to the fact that by possessing the database of the whole area a more adequate model can be created if we also include the characteristics of the neighbourhood around the given cell.

4. CONCLUSIONS

Our aim was to create a model in order to estimate the intensity and spatial distribution of the mean maximum heat island, with the help of urban-surface parameters as well as mathematical-statistical methods. In the course of this work we applied some new parameters which describe urban geometry in three dimensions. Out of these two parameters, the application of spatial compactness as a predictor appeared to be more successful in the model.

As far as we know, such a detailed measurement and analysis of surface geometry for urban climate research have not yet been carried out in our region. Because of the size of the town, as a first step we studied only one sample area representing the whole structure of the town, but not the entire town. The selection of the sample area was carried out by the method of stratified sampling, of which the basic index-number is the built-up ratio, the most important urban surface parameter. After the establishment of the six layers we chose one-third of each class, paying special attention to the spatial distribution. Moreover, it was an important task to represent those areas where large horizontal temperature gradient can be detected in time of development of the *UHI* maximum.

In the course of our research it was possible to measure the spatial data of 11,000 buildings with great accuracy and thus, we could perform a more complex analysis of the connection between urban geometry and the heat island. The compactness, similarly to the predictors describing the urban surface, strongly correlates with the *UHI* intensity; in addition, it became clear that it provided an even stronger connection than the internationally-accepted *SVF* parameter.

With the application of the stepwise multiple linear regression model we could determine coefficients showing in what extent each parameter takes part in the creation of the annual mean *UHI* intensity. Using this model-equation, the absolute deviations – calculated for an independent one-year period – of the spatial extension of the generated heat island remained under 0.5°C almost in the entire investigated area of the town, which is an appropriate result. The structure of the calculated heat island in its characteristic features also showed clear similarities to the real conditions.

In this study, one part of the current results in the urban climatology research of Szeged is discussed in detail. The next step of our project is to finish the 3D urban geometry survey, which helps us to provide a more exact model of the *UHI*. Moreover, in this model it becomes also possible to take the neighbouring cells into consideration. Our further aim is to extend the model towards other towns with favourable conditions for urban climate research (e.g. Debrecen, Hungary). Thus it would become possible to build up such a general model that would enable us to calculate the spatial extent of the mean maximum heat island, practically without any temperature measurements, merely with the application of urban surface parameters. Such data are available for more and more settlements. Therefore, by estimating heat island structure and intensity, which have significant

influence on energy consumption and comfort sensation, such a simple model can provide an adequate help in urban planning.

Acknowledgement – This research was supported by the grant of the Hungarian Scientific Research Fund (OTKA T/049573).

REFERENCES

Balázs, B., Gál, T., Zboray, Z. and Sümeghy, Z., 2005: Modelling the maximum development of urban heat island with the application of GIS based surface parameters in Szeged (part 1): temperature, surveying and geoinformatical measurements methods. *Acta Climatologica et Chorologica Univ. Szegediensis 38-39* (this issue), 5-16.

Bottyán, Z. and Unger, J., 2003: A multiple linear statistical model for estimating the mean maximum urban heat island. *Theor. Appl. Climatol. 75*, 233-243.

Bottyán, Z., Balázs, B., Gál, T. and Zboray, Z., 2003: A statistical approach for estimating mean maximum urban temperature excess. *Acta Climatologica Univ. Szegediensis 36-37*, 17-26.

Cohran, W.G., 1963: *Sampling Techniques (2nd ed.)*. John Wiley & Sons, New York.

Dalenius, T. and Hodges, J.L., 1959: Minimum variance stratification. *J. Am. Statistical Assoc. 54*, 88-101.

Oke, T.R., 1981. Canyon geometry and the nocturnal urban heat island: comparison of scale model and field observations. *J. Climatol.*, 237-254.

Péczely Gy., 1979: *Klimatológia (Climatology)*. Tankönyvkiadó, Budapest.

Sümeghy, Z. and Unger, J., 2003: Classification of the urban heat island patterns. *Acta Climatologica Univ. Szegediensis 36-37*, 93-100.

Unger, J., 2004: Intra-urban relationship between surface geometry and urban heat island: review and new approach. *Climate Research 27*, 253-264.

Unger, J., Bottyán, Z., Sümeghy, Z. and Gulyás, Á., 2000: Urban heat island development affected by urban surface factors. *Időjárás 104*, 253-268.

Unger, J., Bottyán, Z., Sümeghy, Z. and Gulyás, Á., 2004: Connection between urban heat island and surface parameters: measurements and modelling. *Időjárás 108*, 173-194.