

REGRESSION ANALYSIS: A METHOD FOR EXTRACTING INFORMATION FROM SATELLITE IMAGES

Michael Wüthrich

Department of Meteorology and Climatology - University of Basel
Spalenring 145, CH-4055 Basel, Switzerland
Commission III

Abstract

Regression analysis is used for enhancing spatial resolution of thermal information of a LANDSAT-TM 5 image. This information layer can be used as an input for an urban area classification, which then can be used for extracting a functional link between the brightness temperatures and urban structures, again using the regression model. Finally, a simulation of a possible change of urban structures and its impact on brightness temperatures (as one important bioclimatological factor) is carried out.

Keywords: Regression Analysis, Classification, Image Analysis, LANDSAT, Thermal, Simulation

INTRODUCTION

While experimenting with different classification approaches in a selected area¹ around Basel (Switzerland) in the REKLIP²-project (area $\approx 35000\text{km}^2$), the possibilities of including the thermal band 6 of LANDSAT-TM (see table 1) for classification purposes was analysed. One of the problems using this band, is the different resolution of thermal information (it contains 16 pixels of the other bands resolution; see table 1), which either reduces the resolution of the classification or reduces the use of the thermal information. As the major interest of the classification was to differ the urban areas, a lot of the separating information can definitely be found in the thermal band.

MULTIPLE LINEAR REGRESSION ANALYSIS

The assumption made by a linear regression model is that the value for one response variable (y) can be

¹ $\approx 576\text{km}^2$ with the city of Basel in its centre (located in the Upper Rhine Valley with the Jura in the south and the Black Forest in the north)

²REgionales KLimaProjekt: a trinational project between Germany, France and Switzerland

explained by a simple linear function of the k independent variables (x_1, x_2, \dots, x_k):

$$y = c_0 + \sum_{i=1}^k c_i x_i \quad (1)$$

The characteristic statistics can be defined and all data then standardised to avoid scale dependencies. The regression values can be received by inverting the correlation coefficient matrix:

$$\hat{R} \vec{c}^* = \vec{g} \iff \vec{c}^* = \hat{R}^{-1} \vec{g} \quad (2)$$

\hat{R} $k * k$ partial correlation coefficient matrix of the k independent variables

\vec{g} vektor with the correlation coefficients between each independent variable x_i and the dependent variable y

\vec{c}^* vektor with the partial regression coefficients of the standardised data

By destandardising the regression coefficients of the standardised data the regression results of the raw-data can be received:

$$c_i = c_i^* \frac{\sigma_y}{\sigma_i} \quad \text{and} \quad c_0 = \bar{y} - \sum_{i=1}^k c_i \bar{x}_i \quad (3)$$

Using the ordinary least square procedure to determine the regression coefficients brings along restrictions which are encapsulated in the Gauss-Markov theorem. The multiple linear correlation coefficient can be calculated as:

$$r^2 = \sum_{i=1}^k c_i r_{iy} \quad (4)$$

ENHANCING THE SPATIAL RESOLUTION OF THE THERMAL INFORMATION OF A LANDSAT-TM 5 IMAGE

Using additional geographical information, Scherer and Parlow (1990) developed a method to enhance spatial resolution of NOAA-images. Here the aim was

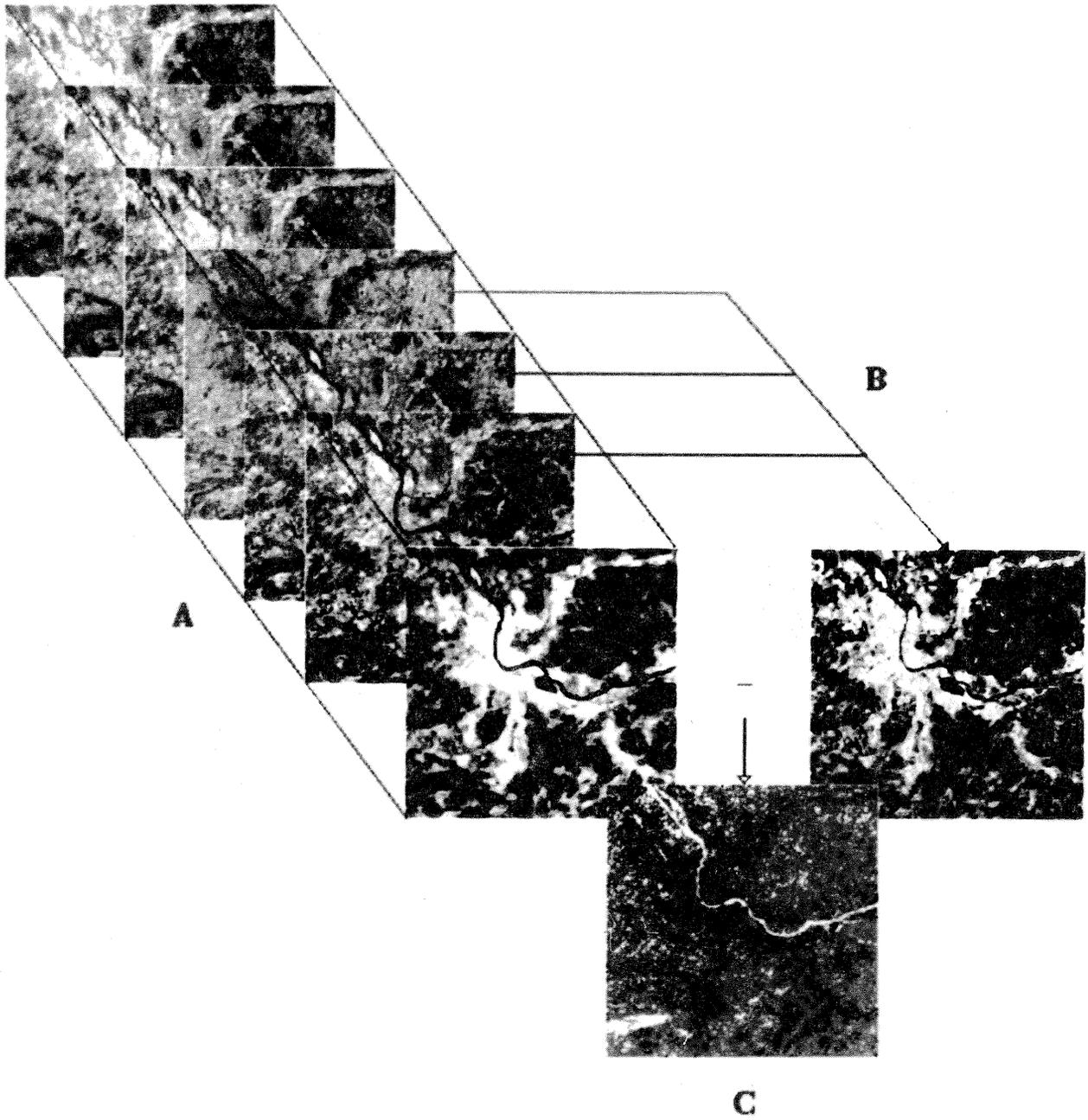


Figure 1: Multiple Regression: (A) Regression Analysis. (B) Using the results to recalculate the dependent variable. (C) Residual between original and recalculated dependent variable (grey: within $\pm\sigma$; black/white: below/above)

Band #	Spectral bands μm	IFOV (m)
1	0.45 - 0.52 (blue)	30 x 30
2	0.52 - 0.60 (green)	30 x 30
3	0.63 - 0.69 (red)	30 x 30
4	0.76 - 0.90 (near IR)	30 x 30
5	1.55 - 1.75 (mid IR)	30 x 30
7	2.08 - 2.35 (mid IR)	30 x 30
6	10.4 - 12.5 (thermal)	120 x 120

Table 1: Characteristics of LANDSAT-TM

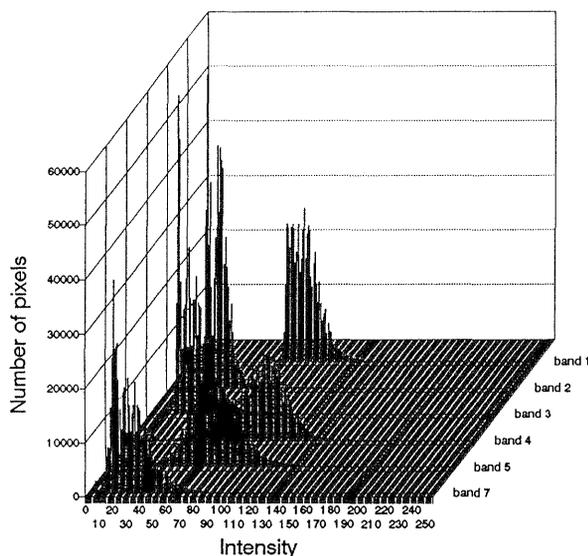


Figure 2: Histogramm: independent datalayers

to enhance thermal resolution of band 6 without any additional data besides the satellites measured bands.

In the above mentioned area around Basel the original data³ (for spatial resolution and spectral bands see table 1) was geocoded⁴. Band 1 to 5 and 7 as independent and band 6 as dependent variables were chosen for the multiple regression analysis. As the resolution of the independent bands is four times higher than the thermal band 6, the former had to be adapted to band 6's resolution. As the satellite's sensor does the same operation (though physically) for band 6, a 4*4-arithmetic mean filter was used and then reduced taking every fourth line and column. The resulting datalayers were then used for the *multiple regression analysis* and are displayed in figure 1 (A) with band 1 in the back and band 6 in the front. The histograms (see figure 2) of the independent variables show almost a normal distribution for each band. Doing so for other areas (such as in polar regions) makes it necessary to use alternative sampling distributions

³LANDSAT-TM of 07.07.1984; 0000 cloud coverage)

⁴Coordinates of the Schweizerische Landestopographie

datalayer	regression coefficient
band 1	not significant
band 2	not significant
band 3	not significant
band 4	-0.122E+00
band 5	0.109E+00
band 7	0.208E+00
regression constant	0.121E+03
mult. correlation coeff.	0.846E+00
explained variance	71.65%
order (band)	7, 4, 5
increase of variance	64.7%,70.4%,71.7%

Table 2: Regression results

(due to the multimodal distributions). After calculating the correlation coefficient matrix and the covariance matrix for the multilayer, the regression was done (results are displayed in table 2), and then, using these results, the synthetic band 6 was produced (see figure 1 (B)) according to equation 1. Although the dependencies between the bands are most probably anything but linear, in this approach, the three bands in the visible spectrum are of no significance for the explanation of the thermal band (table 2), whereas the remaining three bands of the infrared can explain 72% of the variance of the dependent variable with a multiple correlation coefficient of 0.846. Comparison of the original band 6 (figure 1 (A); front) to the synthetically calculated band 6 (figure 1 (B)) shows that the structural information is represented by the recalculated dataset. The *cutting* of the extrema (the river Rhine gets brighter and the city of Basel is gets darker) is a result of the regression procedure (see above).

The linear regression was calculated on a spatially reduced dataset, but the use of the regression coefficients in equation 1 can be applied to the original band 4, 5 and 7 in their 30m*30m resolution. *The resulting synthetic band 6 has now a 30m*30m resolution and can be used for classification purposes.*

RESIDUAL BETWEEN ORIGINAL AND SYNTHETIC BAND 6

For image analysis, it is now interesting to have a look at the areas which *cannot be explained* by this linear approach and have *typical deviations*. So the residuals between the synthetically recalculated band and the original band 6 were calculated according to equation 5 for each pixel.

$$P_{xy} = P_{synthetic_{xy}} - P_{original_{xy}} + 128 \quad (5)$$

The resulting datalayer is displayed in figure 1 (C) with a resolution of 120m*120m. There are only three differentiations made in the display of the residual:

1. grey: within $\pm\sigma$.
2. black: below $mean - \sigma$. Track systems and high density built-up areas and part of the forest areas (caused by the sun exposed steep slopes).
3. white: above $mean + \sigma$. Water, part of the forest areas (caused by the not-sun exposed steep slopes).

Doing the same operation for the high resolution dataset (using band 6 original 30m*30m (that means always 16 equal pixels) and the above described resulting synthetic band 6 in 30m*30m) results in a much more differentiated residual with typical deviations for individual classes (example: the runways of the airport have typical deviations). This newly created datalayer was added to the original bands and a maximum likelihood classification led to excellent results within the focused urban study area⁵ (for classification and a discussion of its results see Wüthrich, 1991). Amongst the rural classes, it was possible to separate 16 urban classes.⁶

BRIGHTNESS TEMPERATURE AND URBAN STRUCTURES

The thermal pattern of a landscape is strongly influenced by the surface characteristics, especially in urban agglomerations (Oke, 1987). The regression analysis represents a method to quantify the influences of different built-up areas on the spatial distribution of radiation temperatures. In this context the following topics have to be mentioned:

- The longwave emission/radiation temperature is of fundamental importance for the energy balance, especially during autochthone weather conditions. The net radiation is described in equation 6.

$$E^* = (S + E_d \downarrow) * (1 - \rho) + E_l \downarrow - E_l \uparrow \quad (6)$$

S direct solar irradiance

$E_d \downarrow$ diffuse solar irradiance

⁵for the classification a further dataset was created: a variance filtered vegetation index (NDVI) in order to get the differences of uniform areas and spatially rapidly changing areas

⁶The accuracy of the result was an average 80% of correctly classified pixels within the training areas. The area within Switzerland was superimposed with the communal boundaries to get classification results for each community. So these results could be compared to conventionally achieved data (such as A: planimetered on topographical maps using aerial photographs and B: official areastatistiks of Switzerland)

ρ	Albedo
$E_l \downarrow$	atmospheric counter radiation
$E_l \uparrow$	terrestrial longwave emission
E^*	net radiation

The radiation temperature T_R is directly connected with the longwave emission through the law of Stefan-Boltzmann (equation 7) .

$$E_l \uparrow = \sigma T_R^4 \quad (7)$$

$E_l \uparrow$ terrestrial longwave emission

σ Stefan-Boltzmann-constant
($5.6697 * 10^{-8} W m^{-2} K^{-4}$)

T_R radiation temperature

- The quantification of the landuse influences improves the understanding of the energetic processes between surface and atmosphere.
- This quantification is a prerequisite to simulate synthetic thermal images under the conditions of a modified landuse (e.g. as a consequence of urban planning)⁷.

ANALYSING THE INFLUENCE OF THE LANDUSE ON THE RADIATION TEMPERATURE

Using a multiple linear regression analysis, which explains the radiation temperature T_R as a linear combination of the percentages of the different landuse classes i ($i = 1$ to n classes) within the pixel $P_{i,xy}$, multiplied with a regression coefficient and summed up with a regression constant. The spatial resolution of the classification data (30m * 30m) and the thermal data (120m * 120m) differ, so the percentage of the landuse classes is calculated from 16 pixel (4*4), which build up one thermal pixel. As the satellite's sensor measures an energie flux, the values of band 6 had to be transformed to brightness temperatures using the method of Schott and Volchok (1985). The regression results for the urban area are listed in table 3. The explained variance reached 88.4% . Using these results, it was possible to recalculate the radiation temperatures inside the study area. The original data received from band 6 is displayed in figure 3 (A) for a selected frame. The recalculated/simulated radiation temperatures are placed on its right side in figure 3 (B). Calculating the difference of the original data $P_{orig,xy}$ minus the synthetic data $P_{syn,xy}$ for each pixel indicates that 67% of the resulting $P_{diff,xy}$ are within the range of $\pm 0.9K$ (with $max_{pos} = +3.6K$ and $max_{neg} = -5.2K$).

⁷The influence of vegetation changes on the net radiation has been quantified (Parlow and Scherer, 1991) for an area in Swedish Lappland

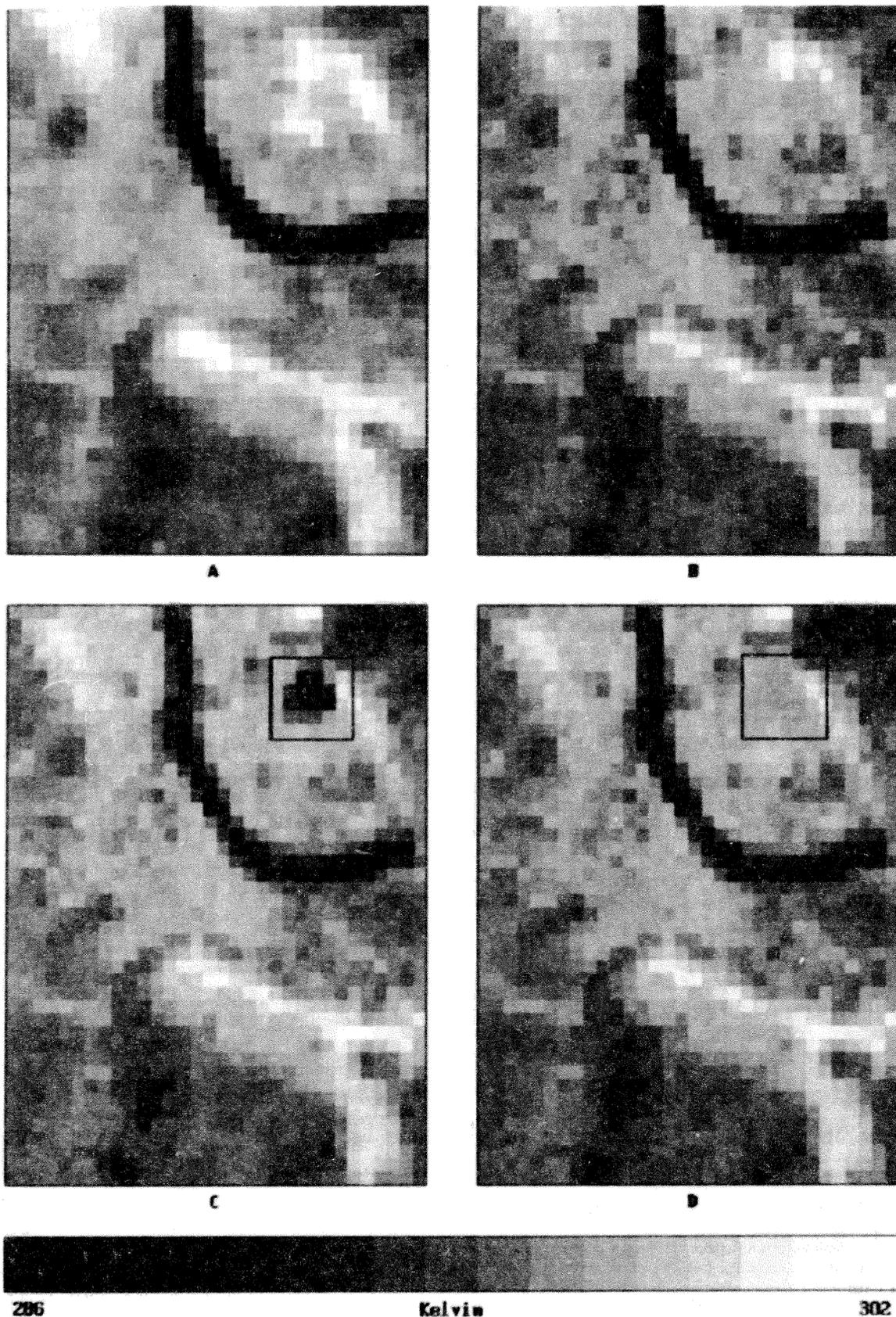


Figure 3: Radiation temperature. Pixelresolution 120m*120m. (A): LANDSAT-TM band 6 original. (B): Recalculated using the multiple linear regression method. (C): Simulated new landuse *park* within the rectangle (D): Simulated new landuse *CBD* within the rectangle

Medium density	0.0171 K per %
High density	0.0422 K per %
CBD	0.0332 K per %
Industrial	0.0542 K per %
Track system	0.0722 K per %
Riverside	-0.0365 K per %
Road	0.0431 K per %
Water	-0.0670 K per %
Forest	-0.0367 K per %
Park	-0.0379 K per %
Regression constant	293.7553 K
Explained variance	88.4 %

Table 3: Regression results (urban area)

The Rhine river in both images appears as the structuring element with the lowest radiation temperatures, which separates the upper right quadrant from the other parts. The bright (high radiation temperatures) areas from the center to the right and in the middle of the upper right quadrant are affected by track systems. An industrial area causing high radiation temperatures is located in the bottom right corner.

SIMULATION OF THE INFLUENCE OF LANDUSE CHANGES

An area of approximately 0.26km^2 in the upper right quadrant (rectangle) was selected for a simulation of change of landuse and its impact on the radiation temperatures. The area is currently used as a track system. Two possible landuses were assumed: park and CBD (Central Business District). The different landuses were implemented into the classification and then using the regression results from table 3, radiation temperatures were recalculated. The results are displayed as images in figure 3 (C) and (D). In the left image (C), the change in radiation temperature for a possible use as a park to the existing landuse is an average -6.0K . Further information is listed in table 4. With a possible use as a high density CBD-area the average radiation temperature differs by -1.5K to the present use. Comparison of the two possible future landuses in the selected area results in an average difference of -4.5K in the park area.

CONCLUSION

Thought as an analysing method for linear dependencies, the regression model and its application mutated as a possibility to reveal latent information hidden in the measured satellite data on one hand and of simulating radiation temperatures under different

	track system - park [K]	track system - CBD [K]	park - CBD [K]
mean	+6.0	+1.5	-4.5
σ	3.5	1.2	2.3
min	+0.4	+0.0	-0.4
max	+10.7	+3.6	-7.1

Table 4: Different landuses and their impact on radiation temperatures (differences)

landuse conditions on the other hand. Using all the bands of LANDSAT-TM, band 6's resolution can be enhanced to $30\text{m}\times 30\text{m}$. Using a landuse classification, it is possible to recalculate/simulate band 6 very precisely (figure 3 (A) and (B)) in a $120\text{m}\times 120\text{m}$ resolution. The application of this method is to get into climate modelling on a regional or local scale. Communal planning authorities normally have detailed plans which imply a modification of the existing landuse, e.g. definition of new industrial areas etc.. This modified future landuse can be integrated in the existing landuse dataset, and then a further simulation of radiation temperatures with the modified landuse can be carried out (figure 3 (C) and (D)). Alternative planning variants can be simulated and then compared to each other. So the climatological effect of planned landuse modification can be quantified very detailed in a local scale using a simple linear regression model.

REFERENCES

- Oke, T.R., 1987. *Boundary Layer Climates*. 2nd ed., Cambridge.
- Parlow, E., Scherer D., 1991. *Auswirkungen von Vegetationsveränderungen auf den Strahlungshaushalt*. Regio Basiliensis. Heft 32/1. Basel
- Scherer D., Parlow E., 1990. *Calculation of synthetic thermal images with enhanced spatial resolution from NOAA-AVHRR satellite images*. In: *Proceedings of EARSeL-Symposium at Helsinki 1989*. Brüssel-Luxemburg 1990. pp. 448-453.
- Schott, J.R., Volchok, W.J., 1985. *Thematic mapper thermal infrared Calibration*. *Photogrammetric Engineering and Remote Sensing* 51, 9. pp. 1352-1357.
- Wüthrich, M., 1991. *Modelling the thermal structures of Basel (Switzerland): a combined approach using satellite data and GIS-techniques*. In: *Proceedings of the 11th EARSeL symposium, Graz-Austria*, pp. 298-304.