**Objective**: Start from predicting existing official economical statistics on the spatial scale where it exists further extrapolating to smaller/different spatial scales creating economical temperature indexes.

At one hand, we consider available official economical statistics such as GDP, Housing prices, Unemployment rate, Level of higher education, Life duration and Crime rate, available on the province scale.

At the other hand we create a feature space by defining various meaningful characteristics of spatial area performance through activity of it’s businesses, residents and visitors using the bank card spending data. All indicators are normalized between 0 and 100 by fitting an appropriate distribution and taking an inverse cumulative distribution function. Right now we consider the following 22 indicators.

1. ***Density of the spending activity within the area***

Average density of number of transactions made within 1sq.km of the area

1. ***Density of the earnings within the area***

Average density of amount of money earned within 1sq.km of the area

1. ***Average amount of a single transaction earned within the area***

Average amount - ratio between total amount and number of transactions made by all customer within the considered area

1. ***Average activity of the customers living in the area***

Average number of transactions per customer – ratio between total number of transactions made by area residents and number of active residents

1. ***Average transaction performed by customers living within the area***

Average amount - ratio between total amount and number of transactions made by residents of the considered area everywhere in the country

1. ***Out-of-province activity within the area***

Percentage of the number of transactions made within the area by it’s out-of-province visitors

1. ***Foreign activity within the area***

Percentage of the number of transactions made within the area by it’s foreign visitors

1. ***Area’s earning diversity***

Number of top business categories (of 76) enough to cover 80% of the total number of transactions got by the area

1. ***Area’s spending diversity***

Number of top business categories (of 76) enough to cover 80% of the total number of transactions made by the area residents

1. ***Area’s business density***

Number of active businesses within the area per sq.km

1. ***Average business size within an area***

Average earnings of an active business within the area (right now it’s simply a total amount earned divided by the number of active businesses, but worth considering a normalized version with respect to average sizes per category, i.e. Total\_Area\_Earning/ sum\_category number\_of\_businesses(category)\*avg\_size(category))

1. ***Percentage of gas/parking/toll spending of area’s residents***
2. ***Percentage of taxi spending of area’s residents***
3. ***Percentage of public transportation spending of area’s residents***
4. ***Percentage of cafés/restaurants spending of area’s residents***
5. ***Percentage of fastfood spending of area’s residents***
6. ***Percentage of food spending of area’s residents***
7. ***Percentage of recreation spending of area’s residents***
8. ***Percentage of fashion/beauty/jewelry spending of area’s residents***
9. ***Percentage of medical spending of area’s residents***
10. ***Percentage of cultural spending of area’s residents***
11. ***Percentage of travel spending of area’s residents***

Also to be implemented:

1. ***Percentage of area nighttime earnings***

Percentage of area’s number of transactions made between 22:00 and 6:00

1. ***Percentage of area weekend earnings***

Percentage of area’s number of transactions made on Saturday-Sunday

1. ***Percentage of area’s residents nighttime spending***

Percentage of transactions made by people living in the area between 22:00 and 6:00

1. ***Percentage of area’s residents weekend spending***

Percentage of transactions made by people living in the area on Saturday-Sunday

1. ***Area’s business integration***
2. ***Area’s residents local mobility (radius of gyration)***

Mean radius of gyration of customers residing in the area during the day computed for each customer as ((sum\_i |r\_i-r\_mean|^2)/N)^0.5, where r\_i are the vectorized positions of the customer during the day, r\_mean is their average, N - number of positions. In order to exclude unusually big distances, after calculating the mean location for the customer we exclude all r\_i's which are beyond a certain reasonable radius from it (say 100km) and then recalculate mean once again. For the mean radius we compute a weighted average weighting each customer daily activities by N-1 (i.e. customers with just 1 transactions during this day are not included as their radius is 0), while the more activity the higher is the weight.

1. ***Area’s residents spending in expensive locations***

Percentage of the total transaction of area’s residents made in the businesses, which average transaction amount is twice (or other threshold) above average for the corresponding business category

1. ***Area’s price level***

Average amount of purchase in the area vs averages for the corresponding business categories: Total\_area\_earnings/sum\_business transactions(business)\*avg\_purchase(category(business))

1. ***Level of prices paid by area’s residents***

Average amount of purchase of area residents with respect to the corresponding business category averages: Total\_area\_spending/sum\_customer,category transactions(customer,category)\*avg\_purchase(category)

1. ***Area’s customer loyalty***

Average number of transactions per business made by area’s residents, i.e. total number of transactions made by them divided by a cumulative number of customer-business relations (where a customer is an area resident and the business is any business countrywide/or within the same province just to exclude occasional visits while travelling). Can be also normalized with respect to average loyalty per category.

Of course many of the indicators are strongly correlated with each other, so the first step of the approach is performing dimensionality reduction through a standard principle component analysis. It gives principle components with the following impact on the total variation of our indicators

**Table 1.** Results of principle component analysis

|  |  |  |
| --- | --- | --- |
| Component | % of variation added by component | Cumulative % of variation of all the previous components |
| 1 | 82.6410 | 82.6410 |
| 2 | 5.9418 | 88.5829 |
| 3 | 3.9438 | 92.5267 |
| 4 | 1.3042 | 93.8309 |
| 5 | 1.1219 | 94.9527 |
| 6 | 0.9293 | 95.8820 |
| 7 | 0.7391 | 96.6211 |
| 8 | 0.6197 | 97.2409 |
| 9 | 0.4983 | 97.7392 |
| 10 | 0.4691 | 98.2083 |
| 11 | 0.3336 | 98.5419 |
| 12 | 0.2817 | 98.8236 |
| 13 | 0.2586 | 99.0822 |
| 14 | 0.2280 | 99.3102 |
| 15 | 0.1866 | 99.4968 |
| 16 | 0.1713 | 99.6681 |
| 17 | 0.1554 | 99.8235 |
| 18 | 0.0821 | 99.9056 |
| 19 | 0.0580 | 99.9636 |
| 20 | 0.0280 | 99.9916 |
| 21 | 0.0083 | 99.9999 |
| 22 | 0.0001 | 100.0000 |

So we see that only first 6 components have an impact of at least around 1% and in total they already cover over 95% of the entire variation.

The above components are linear combinations of the normalized values of our 22 indicators and could be thought as activity temperatures representing 6 different aspects of commercial activity (as distinctive as possible). Table 2 reports the weights (normalized so that sum of absolute values of the weights is 1) with which our indicators are included into those temperature indexes.

**Table 2.** Structure of the six detected commercial temperatures with the weights of indicators linearly included into them; major positive/negative weights are highlighted by green/red.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Indicator\ Component | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 0.0429 | -0.0091 | 0.0008 | 0.0634 | 0.0097 | -0.0199 |
| 2 | 0.0428 | -0.0107 | 0.0011 | 0.0625 | 0.0102 | -0.0205 |
| 3 | 0.0479 | -0.0812 | -0.0651 | -0.0668 | 0.0523 | -0.0372 |
| 4 | 0.0476 | 0.0914 | 0.0132 | 0.0339 | -0.0310 | -0.0373 |
| 5 | 0.0479 | -0.0878 | -0.0678 | -0.0419 | 0.0232 | -0.0487 |
| 6 | 0.0439 | -0.0977 | 0.0373 | 0.0470 | -0.0826 | 0.0163 |
| 7 | 0.0435 | 0.0449 | 0.0427 | -0.0770 | 0.1084 | 0.1197 |
| 8 | 0.0500 | 0.0689 | -0.0614 | -0.0246 | -0.0182 | -0.0175 |
| 9 | 0.0506 | 0.0603 | -0.0670 | 0.0148 | -0.0175 | -0.0430 |
| 10 | 0.0436 | 0.0013 | -0.0135 | 0.0555 | 0.0140 | -0.0231 |
| 11 | 0.0445 | -0.0066 | -0.0273 | 0.0735 | -0.0221 | -0.0411 |
| 12 | 0.0439 | -0.0749 | 0.1632 | -0.0537 | -0.0590 | 0.0196 |
| 13 | 0.0442 | 0.0280 | -0.0084 | -0.0413 | -0.0650 | 0.0079 |
| 14 | 0.0426 | -0.0216 | 0.0155 | -0.0187 | 0.0187 | 0.0230 |
| 15 | 0.0467 | 0.0545 | -0.0326 | -0.0979 | 0.0343 | -0.0751 |
| 16 | 0.0446 | 0.0518 | 0.0416 | -0.0561 | -0.1421 | -0.0446 |
| 17 | 0.0449 | 0.0335 | 0.1810 | 0.0281 | 0.0691 | -0.0108 |
| 18 | 0.0447 | 0.0038 | -0.0098 | -0.0087 | 0.0384 | -0.0097 |
| 19 | 0.0442 | -0.0824 | -0.0213 | 0.0336 | 0.0513 | -0.0707 |
| 20 | 0.0475 | 0.0527 | -0.0149 | 0.0922 | 0.0347 | 0.0928 |
| 21 | 0.0437 | -0.0050 | 0.0227 | -0.0000 | 0.0375 | 0.0150 |
| 22 | 0.0478 | -0.0320 | -0.0919 | -0.0085 | -0.0607 | 0.2066 |

So we see that first temperature simply includes all the indicators with the relatively same weight, second – is primarily defined by activity (avg. transaction amount in the area and by it’s residents, activity of residents and out-of-province activity within the area) together with gas and fashion spending percentage, third – by gas, food and travel spending, fourth – by business size, café and medical spending, fifth – by out-of-province and foreign activity within the area as well as fastfood activity of it’s residents, and sixth- foreign activity as well as café, medical and travel spending of residents.

Focusing on those 6 first principle components, we find the following individual correlations (corr coeff, %) between them and the 6 official statistic parameters (also normalized) on the principle scale.

**Table 3.** Individual correlations between 6 leading components and 6 statistical parameters

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Component | GDP | Housing | Unemployment | Education | Crime | Life |
| 1 | -47.1409 | -0.1102 | -4.4827 | -45.9581 | 35.8273 | 1.1286 |
| 2 | -7.7109 | 68.2208 | -63.8609 | -0.4049 | -35.7114 | -42.9780 |
| 3 | -56.7172 | 31.6483 | -35.7058 | -64.6634 | 24.2448 | -3.7685 |
| 4 | -41.7675 | 33.9962 | -36.2206 | -34.1211 | -63.1003 | -44.0191 |
| 5 | -4.9377 | 6.5361 | -8.0944 | -33.6917 | -24.5502 | -34.8659 |
| 6 | 9.0829 | 2.7000 | -2.4917 | 8.7373 | 6.9622 | 0.8916 |

So as we see component 1 is strongly correlated with GDP and Education, while component 2 – with Housing, Unemployment and Life. Component 3 also has a strong correlation with GDP and education, while components 4 – with GDP and Crime and life expectancy. Component 5 is already rather weekly correlated with the official statistics (Education, Crime and Life expectancy), while component 6 already does not show any significant correlations.

Now we try to fit the 6 statistical parameters through the above principle components, training the logistic regression on different sets of provinces and validating the model on the rest in order to get rid of the overfitting. The results averaged over 4 different combinations of training and validation samples are reported below.

**Table 4.** Fitting/cross-validation results with 6 components – cross validation results

|  |  |  |
| --- | --- | --- |
| **Parameter** | **R2 validation** | **R2 training** |
| GDP | 0.4088 | 0.5564 |
| Housing | 0.5588 | 0.7251 |
| Unemployment | 0.5426 | 0.7144 |
| Education | 0.4694 | 0.6810 |
| Crime | 0.1999 | 0.6498 |
| Life | 0.4676 | 0.5713 |

The R-squared coefficient (% of parameter variation explained by the model) is lower on the validation sample compared to the training sample (which hints for some over-fitting) however for all parameters but Crime R2 stays considerable even on the validation set.

So overall one can summarize the findings as some initial proof of feasibility of predicting economical quantities using suggested indicators, however having just a small sample of only 52 provinces where those statistical parameters are defined, limits our possibility to advance further.

**Next steps**

1. Implement remaining indicators.
2. Compute the indicators on the levels of comarcas and municipalities, performing dimensionality reduction and analyzing the structure of principle components for consistency.
3. Aggregate additional statistical data on the levels of provinces, comarcas and municipalities and try to fit it learning the model over the above feature space. Analyze the consistency of the model structure over different spatial scales.