The Geographical Balance of Regional News of Czech TV CT24

Peter Nemec, Jiří Horák

Institute of Geoinformatics, Faculty of Mining and Geology, VSB-Technical University of Ostrava, 17. listopadu 15
70833, Ostrava, Czech Republic
peter.nemec.st@vsb.cz

Institute of Geoinformatics, Faculty of Mining and Geology, VSB-Technical University of Ostrava, 17. listopadu 15
70833, Ostrava, Czech Republic
jiri.horak@vsb.cz

Abstract. The regional channel of Czech TV 24 has been monitored for six months and an analysis of the spatial and time distributions of news has been performed. The RSS channel of this TV station was automated scanned and geocoded which requested to prepare a set of appropriate procedures how to capture, store news and how to process unstructured texts. Results approve existing rough regional balance of the media news reflecting the number of population. Nevertheless some regions demonstrate evidently high and unexpected account e.g. Karlovy Vary region.

Keywords: RSS, GeoRSS, XML, Geocoding, Geoparsing, News

1 Introduction

Media has to provide a balanced service which is one of the basic requirements for any public service. How can we understand the balance? Usually we wonder if information provided by different political subjects is presented according to the voting preferences.
We can ask ourselves a question if the news are also geographically balanced, i.e. equally distributed across the relevant geographical territory (region, country, world). If we recognise some anomalies (clusters of news) we should explore them and search for the reasoning of such phenomenon. Is it done by a higher popularity of such a place? Or is it just a sequence of news from one place related to one big event? It is clear that for studying a balance of news we need to describe and analyse also the distribution during a time period, not only in the geographical area.

From processing point of view the easiest form of monitoring media news is to explore its RSS channels (most of media provide such internet channels). The explanation of RSS can be found in chapter 3.

To study the distribution of news we have decided to monitor web version of CT24 TV. Its RSS channel publishes news 24 hours per day in several sections – economy, home news, world, sport, culture, regional news etc. We have selected regional news for monitoring and evaluating due to the fact that for such news a geographical balance has to be guaranteed.
This section broadcasts news using RSS channel http://www.ct24.cz/rss/regionalni.

2 Principles of geocoding and geoparsing

Geocoding identifies geographical coordinates on the base of analysis of structured location references like postal addresses. The process is based on the comparison of individual sections of the given address with data in a reference layer (set of municipalities, set of streets etc.).
Geoparsing represents a process similar to geocoding.
Geoparsing is the process of assigning geographic identifiers to textual words and phrases that occur in unstructured content, such as "twenty miles north east of Jalalabad". You can also geoparse location references from other forms of media, for example audio content in which a speaker mentions a place (http://en.wikipedia.org/wiki/Geoparsing).
A location information is recorded by natural language in the text. Geoparsing is also able to process ambiguous references – i.e. Petrovice is the name of several municipalities and we need additional information and appropriate process to choose the right alternative (Charvát et al. 2007)

Geocoding and geoparsing of news is applied i.e. from Slovak newspaper SME (http://www.sme.sk/mapa). The process is described by Bella (2008).

The geocoding and geoparsing are utilised for various purposes. I.e. Beaman and Barry (2003) describe the application of these methods to improve streamlining and automate acquisition of biogeographic data.
The multi-step process involves pre-processing text for language, locale or project specific anomalies (e.g. standardising abbreviations), phrase analysis (the description is compartmentalized by punctuation, prepositions, and stop words into separate phrases. Each phrase is analysed independently.), text parsing and pattern matching using regular expressions will involve detecting feature types (e.g., National Park, Island), place names, and their inter-relationships, calculation of geographic offsets (e.g., 2.5 km WNW of …), recording.
Geoparsing of disease alerts (Keller, Brownstein, Freifeld 2008) using gazetteer approach and utilisation of neural networks are applied to improve the georeferencing capability of the HealhMap server (www.healthmap.org).

Our approach combines both geocoding and geoparsing for georeferencing media news in a multi-step simple process.

3 RSS channel

RSS (Really Simple Syndication) is a family of Web feed formats used to publish frequently updated works – such as blog entries, news headlines, audio, and video – in a standardized format. An RSS document (which is called a "feed", "web feed", or "channel") includes full or summarized text, plus metadata such as publishing dates and authorship. A **web feed** (or **news feed**) is a data format used for providing users with frequently updated content http://en.wikipedia.org/wiki/RSS_(file_format).
A RSS document usually contains new headlines and text of news and publish them on a unique URI address.
By its internal form a RSS document is a subset of XML. More about version of RSS can be found in (RSS SPECIFICATIONS: History of RSS, *2008)* or *(*HOLZNER S., Šindelář J., 2007).

4 Reference data

We have selected a level of municipalities to locate news. This decision started with initial estimation of news volume and frequency of individual location. More detailed georeferencing will provide quite rare occurrences in the country and the location will be too detailed, while the location error will substantially increase. The municipal location seems to be also adequate for envisaged application of such a service.
The list of municipalities was obtained from UIR-ZSJ (Registry of area units) from the Czech Statistical Office.
The list had to be processed to obtain the appropriate data for geocoding and geoparsing.
First, duplicate names were deleted. Secondly, any ambiguous characters (parenthesis, numbers) were eliminated (deleted, substituted etc.). After this stage the number of municipalities dropped down from 6249 to 5328, it means approx. 15% of names were deleted.
Due to the changes of word termination according to grammatical cases it is needed to be able to identify name variants in the news.
One of the possible way how to achieve this is to extend the existing list of municipal names with variants given by all potential grammatical cases.
A new PHP script is able to generate all variants based on the lemat (root of the word) and cases termination (i.e.: *Ostrav**a***, *Ostrav**y,** Ostrav**ě,** Ostrav**ou)***. Such possible cases termination are 49.
These variants are added to the list of municipal names. Of course, the extension causes also data redundancy and increased risk of conflicts.
The final list of municipalities was applied for geocoding and geoparsing.

5 Recording and geocoding of media news

The structure of RSS for CT24 has been analysed. Any news are published using a structure element called item. An example of an item is depicted in the upper part of Fig. 1.

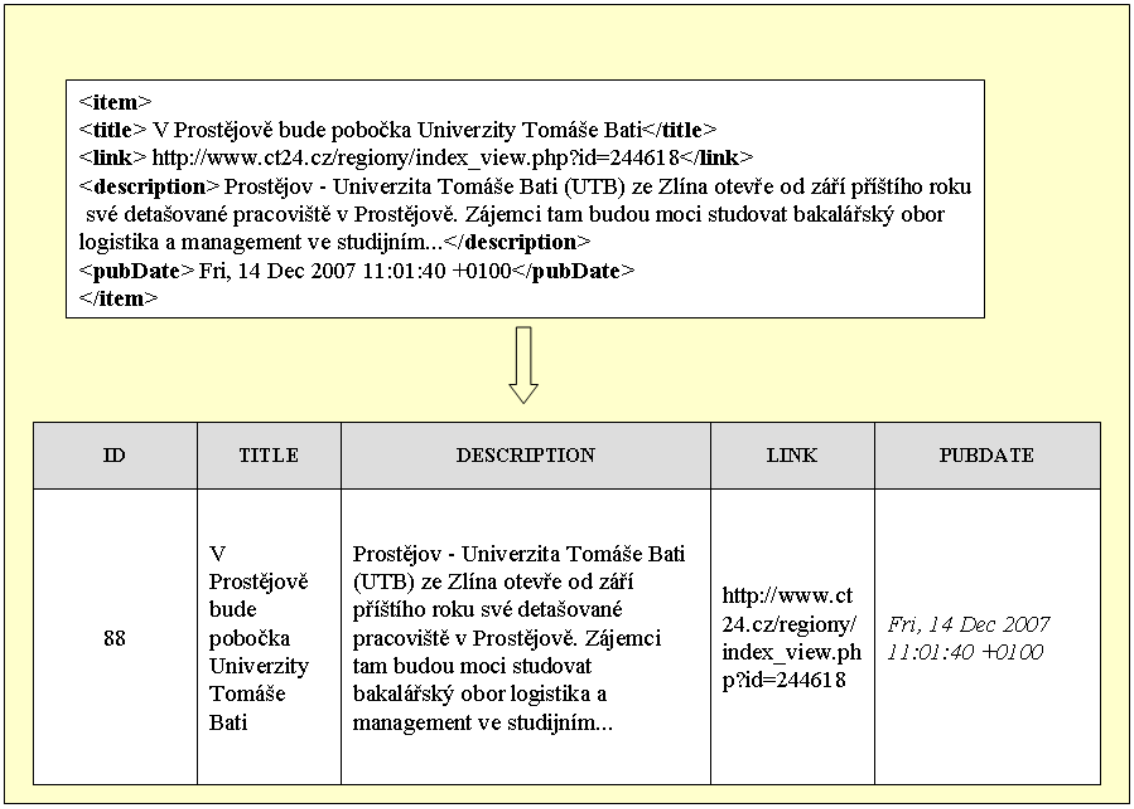

Fig. 1 Content of RSS and recording into a database table

The *<item>* element contains 5 child elements:

<title> - defines the headline of the story;

<link> - defines the hyperlink to the item; (http://www.ct24.cz/regiony/index_view.php?id=244618);

<guid> - defines a unique identifier for the item;

<pubDate> - specifies the day of issue this item;

<description> - contains the abbreviated version of the original story. This is the most essential element, where we can recognise references to locations. In the example above two references have been identified: *Prostějov, Zlín*.

Furthermore the element *<description>* is divided into two separated parts:

domicile, which defines the location of news, and

content, where the story is depicted using unstructured text. Locations may be also included.

Sometimes it is difficult to separate the domicile from the content automatically, because the domicile can be written in different variants.

The content of RSS for CT24 is dynamically generated and changed (approximately every hour) by publishers. Old news are deleted and comming news are added. The second PHP script was created to continuously analyse the RSS channel and to record any new item to a database table. The item is segmented into child elements and inserted into one record (Fig. 1).The script is executed every hour.

Geocoding deals with a reference data and the table of recorded news. The task is to assign locations from the reference data (in this case it is a list of municipal names) to every news. This task is completed by a third PHP script. Any news description is segmented into words, which are compared

to every case termination generated from the reference data table (Fig. 2). Additionally the type of occurrence is recorded (1 for domicile, 2 for text).

ID	TITLE	DESCRIPTION	LINK	PUBDATE
472	Zlínská policie kontrolovala taxikáře	**Zlín** - Policie na Zlínsku při nočním zátahu kontrolovala, jestli taxikáři nejezdí načerno nebo opilí. Hlavním impulzem pro kontroly se stala nedávná tragédie, která se odehrála ve **Vizovicích**, kde jednaosmdesátiletý důchodce zemřel pod koly taxíku, jehož řidič řídil načerno a bez jakéhokoli řidičského oprávnění.	http://www.ct24.cz/regionalni/8864	*Sat, 15 Mar 2008 15:43:00 +0100*
61	U Ostrova unikla z prasklého potrubí nafta do chovného rybníka	**Ostrov** (Karlovarsko) - Chovný rybník u Ostrova na Karlovarsku dnes znečistila nafta, která unikla z prasklého potrubí. Potrubí vede do nedalekého skladu pohonných hmot společnosti Čepro…	http://www.ct24.cz/regiony/index_view.php?id=244104	*Tue, 11 Dec 2007 15:56:00 +0100*

	ID	MUNICIP	CASE_TERMINATIONS
①	5200	Zlín	**Zlín**,Zlína,Zlíně,Zlínem,Zlínu…
②	4882	Vizovice	Vizovice,**Vizovicích**,Vizovicemi,Vizovicěmi,Vizovicími,Vizovicými,Vizovicmi,Vizoviceti,Vizoviceti,Vizovicovi,Vizovici,Vizovicě,Vizovicí,Vizovicám,Vizovicetem,Vizovicětem,Vizovicem,Vizovicém,Vizovicím,Vizovicatům,Vizovicům,Vizovicým,Vizovicého,Vizovicího,Vizovico,Vizovic…
①	3108	Ostrov	**Ostrov**,Ostrovíma,Ostrovýma,Ostrovata,Ostrova,Ostrovová,Ostrová,Ostrovete,Ostrověte,Ostrove,Ostrovové,Ostrové,Ostrovách,Ostrovatech,Ostrovech,Ostrovích,Ostrových,Ostrovami,Ostrovemi,Ostrověmi,Ostrovími,Ostrovými…

Fig. 2 News georeferenced using variants of case terminations

Geocoding is negatively influenced by an ambiguity of municipal names (names are not unique and even more, case terminations generate additional false occurencies), which leads to assign more candidates for one name. For example “Ostrov” is the name of 6 municipalities in the Czech Republic. The last PHP script searches any additional information in the content to improve the rate of selecting the right alternative. The procedure utilises a supplemental information like a name of the district (i.e. Karlovarsko). If no supplemental information is available, the script takes into account other municipalities recognised and recorded for the news (item). It calculates Euclidean distances and selects the candidate closest to other municipalities.

6 Mapping a news’ distribution

Recorded and georeferenced news were deployed to create a statistical map of media new’s occurrences during the selected study time period (19.12.2006 – 18.10.2008) (fig. 3).

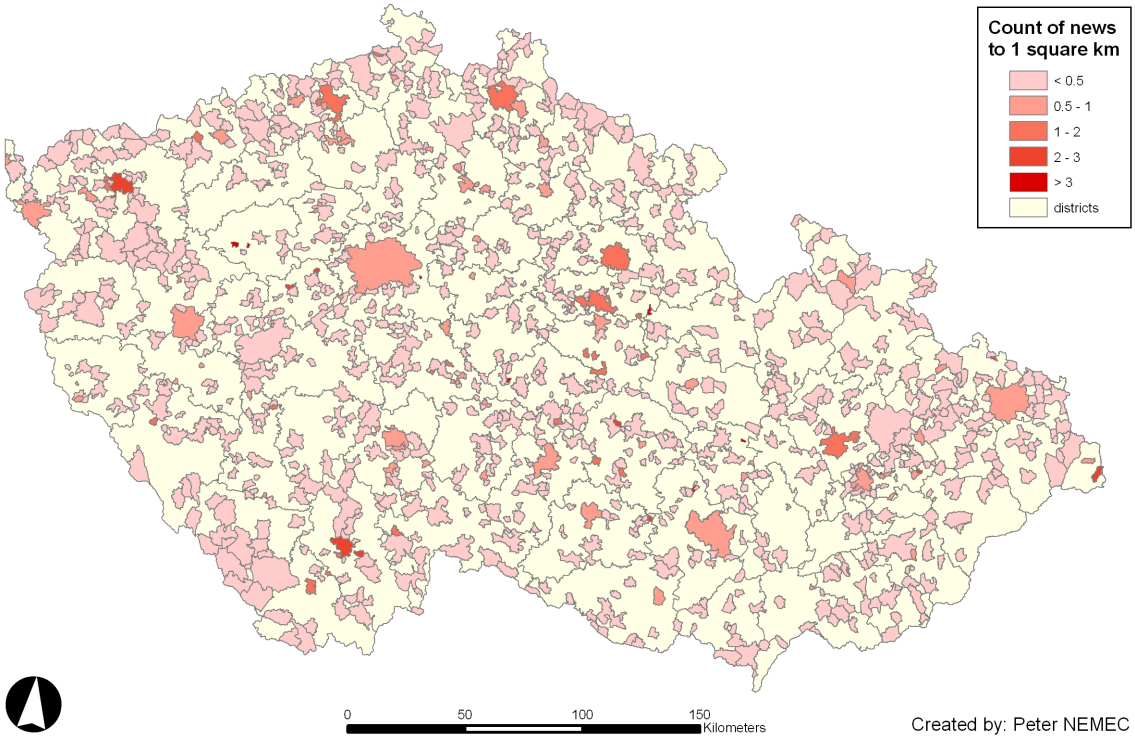

Fig. 3 News density (CT24 Regional News occurred from 19.12.2006 to 18.10.2008)

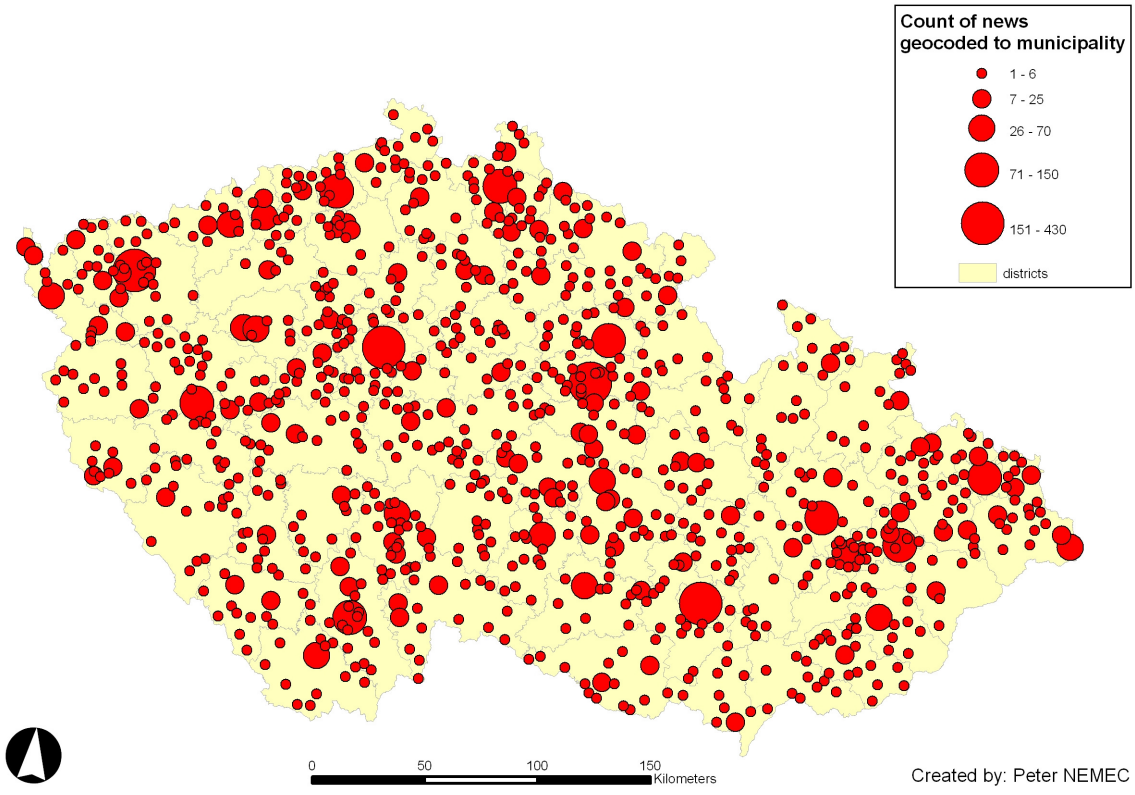

Fig. 4 Municipalities where one or more news of CT24 occurred from 19.12.2006 to 18.10.2008

The map (fig. 3) documents an approximately even distribution of news from the territory of CZ. It confirms the status of „regional news“ of this RSS channel. Nevertheless we can recognised some vacant places where no occurrence was recorded – area around Hodonín and Kyjov, part of Beskydy Mts., central part of Jeseniky Mts., border between Central Bohemia and Southern Bohemia regions, Podbořansko. Some of them represent rural and mountainous areas but it is obvious that a rural aspect is not the only reason, because other rural areas like Krušné Mts. occur frequent in news.
Next map (fig. 4) provides an indication of a relationship between news and population. This relation demonstrates a higher level of correlation and an satisfactory linear regression form (R^2 = 0,74) (fig.5), slightly worse than quadratic form (R^2 = 0,78). Several deviations need to be analysed. Karlovy Vary, Pardubice, České Budějovice, Hradec Králové, Ústí nad Labem, Olomouc and Liberec represent regional centres where the number of news is higher than expected. There are also 2 small municipalities with a high number of attached news due to the missclassification (false geocoding) – Ústí (Přerov district, 88 news for 538 inhabitants), Karlov (Žďár n.S. district, 40 news for 65 inhabitants). We have design the geocoding process to decrease the error of omitting some municipalities which simultaneously increases the error of wrong classification.
The relationship to population is employed in following maps where we use a ratio of news to the number of inhabitants.

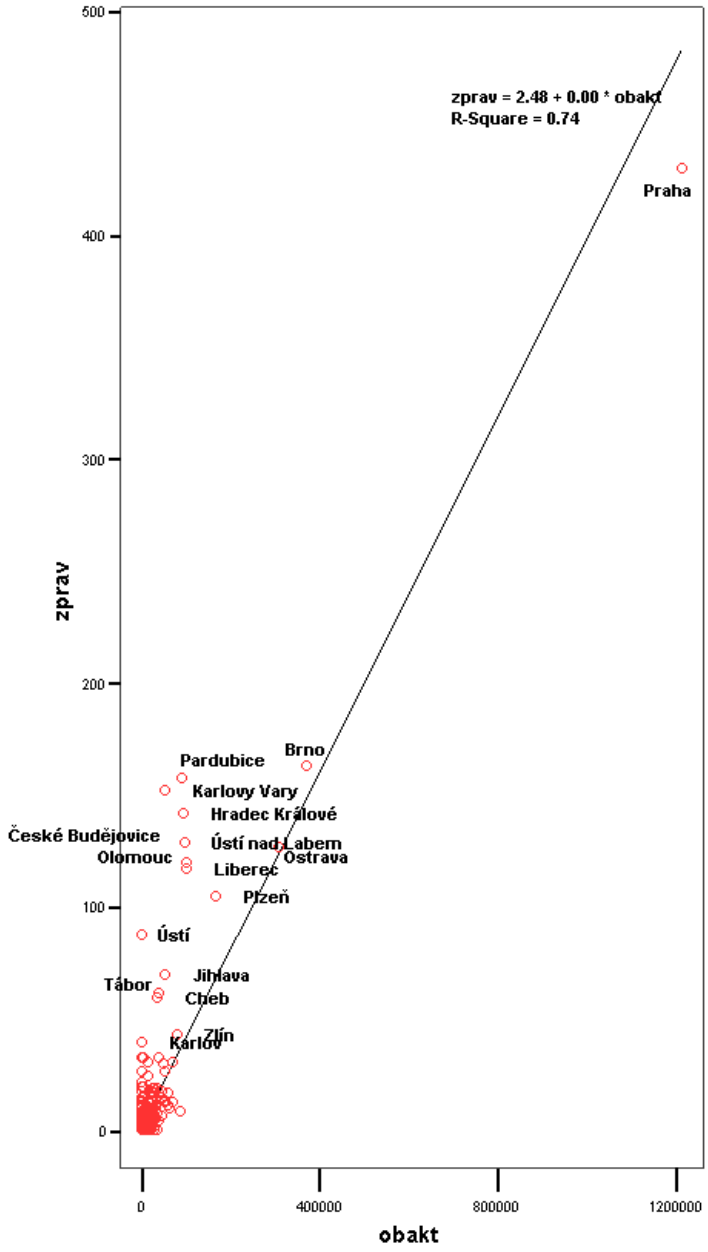

Fig. 5 Count of news to the population of the municipality (1.1.2008)

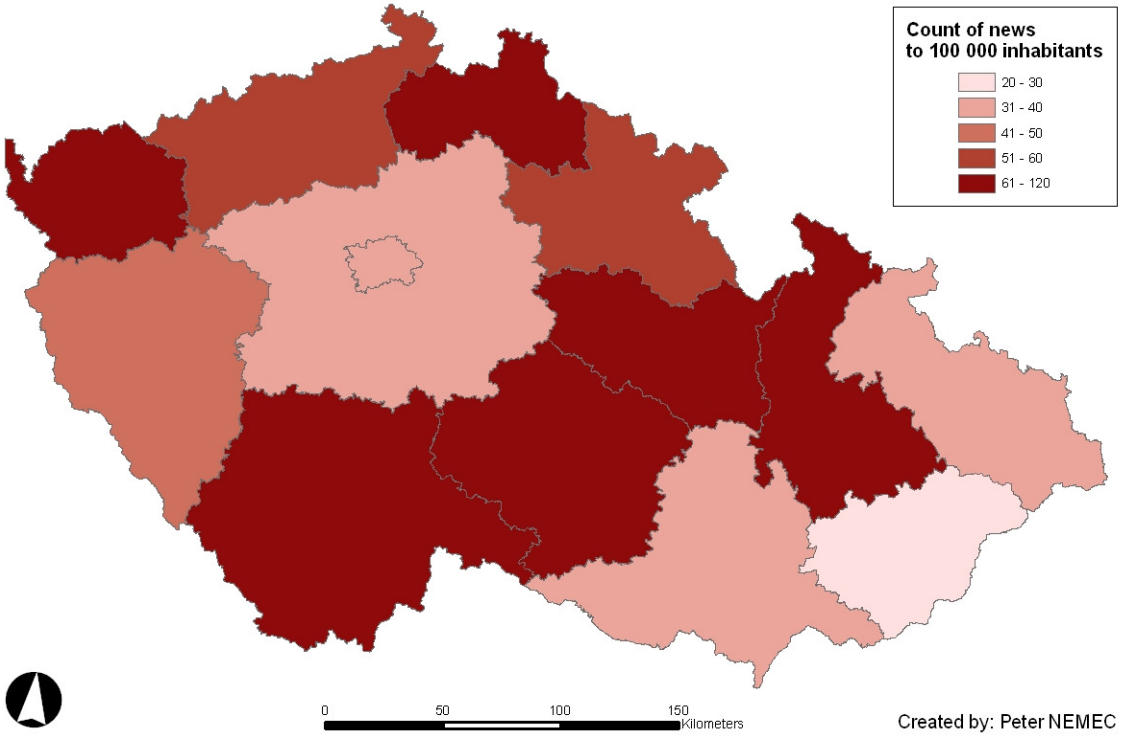

Fig. 6 Count of news to 100 000 inhabitants by region (RSS of CT24, 19.12.2006 – 18.10.2008)

Aggregation of results to the regional units demonstrates a slightly uneven distribution (fig. 6). The highest ratio is carried by Karlovarsky region (app. 120), followed by Jihocesky region (86), Vysocina (78) and Pardubicky region (73).
Surprisingly, the ratio for Prague is relative low (37). It is probable caused by a high population in the capital and an effort to bring more regional than „central“ news.
The lowest ratio is assigned to Zlinsky region (only 22). Reasons may be found in its peripheral and mountainous character.

The more detailed look inside is provided using aggregation to a district level (LAI) (fig.7). Several districts declare a higher ratio of news. Karlovy Vary (17), Usti n.L. (12), Pardubice (14) represents centres of regions, but many other districts cannot be explained this way. There are Cheb (12), Rakovník (13), Tábor (12) and Přerov (13). In the district of Cheb there occurred many news that relates to German municipalities in neighbourhood. After the admission of the Czech Republic to the European Union the count of news increased in municipalities closed to Germany. When we analyse recorded news, the important part of them relates to criminal activities discussed in municipalities close to the border. These activities have often been captured by news. The district of Rakovnik includes several municipalities with common names which can be easily found in any text because they depict common words like “reason” (Příčina), “small town” (Městečko) or frequent first personal name (Václavy).

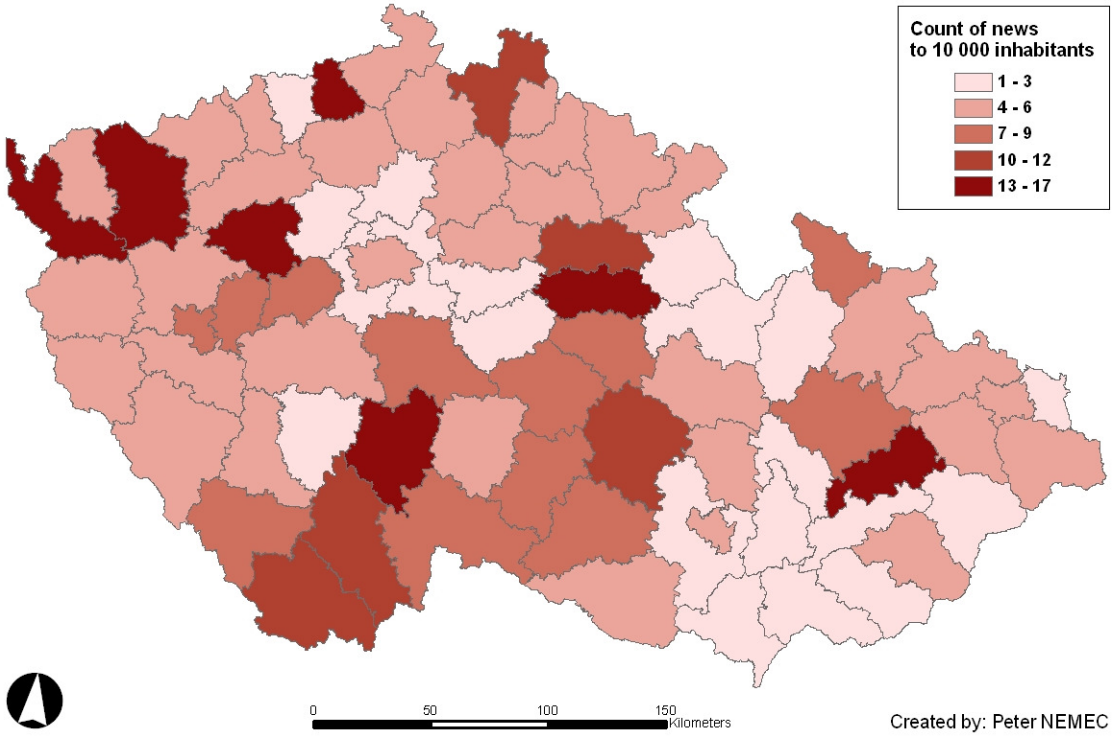

Fig. 7 Count of news to 10 000 inhabitants by district (RSS of CT24, 19.12.2006 – 18.10.2008)

7 Temporal distribution

High geographical clusters of news can be evoked in some cases by an impact of some important event. To study such phenomena we need to measure of time clustering for news.

The NNI (nearest neighbour index) test evaluates distributions of distances (Bailey, Gatrell 1995, Horak 2006). It is based on the average nearest neighbour distance between all events. It compares expected distances to observed distances.
In this case we develop a Temporal Nearest Neighbour Index (TNNI) to evaluate a temporal distribution of our phenomena (events). It compares R_o to R_e where:

R_o – the average of 50% shortest temporal distances between the event and its closest neighbour,
R_e – the estimated average temporal distance between all events.

Results can be interpreted according to a position between following limits:

TNNI = 0 ... clustered sample
TNNI = 1 ... uniform sample

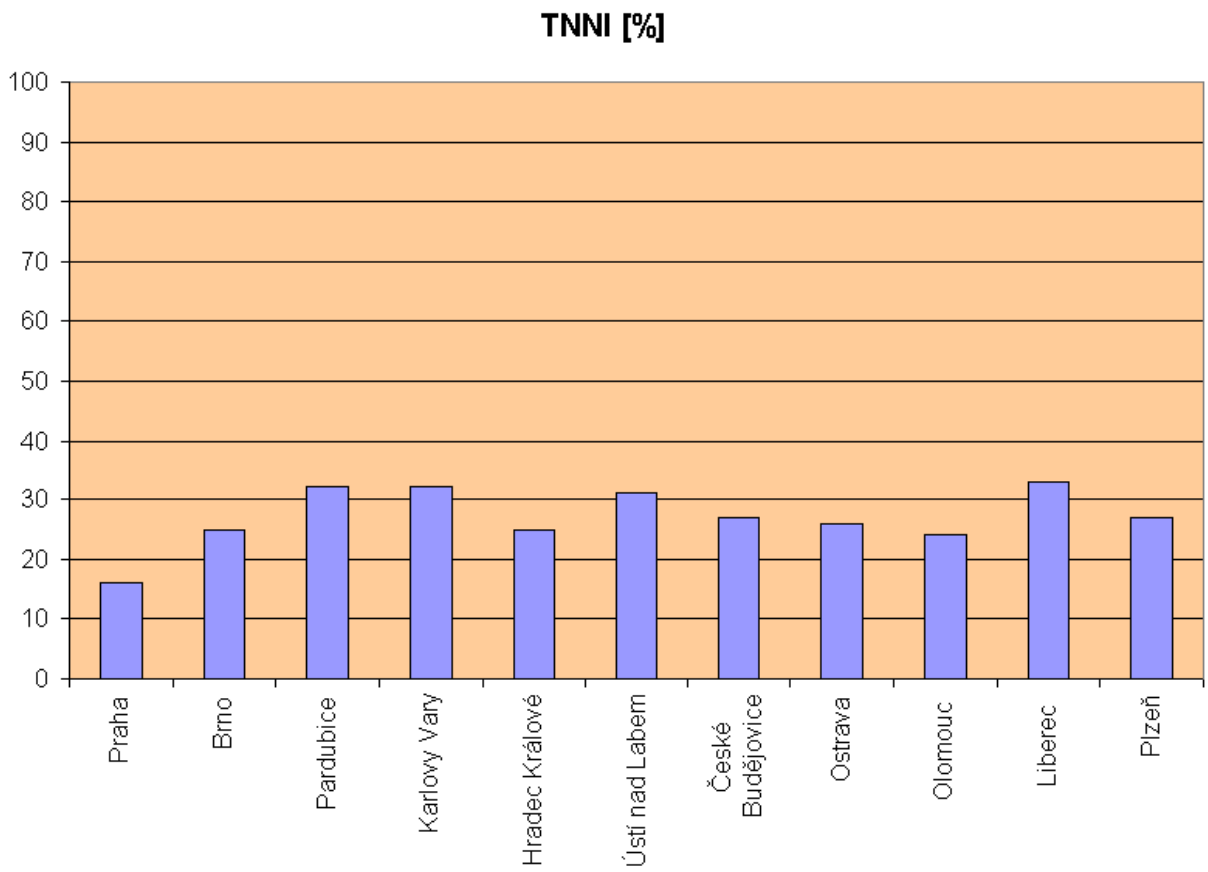

Fig. 8 TNNI (%) for selected municipalities, where 100 and more news occurred (19.12.2006 – 18.10.2008)

The TNNI (fig. 8) ranges between 16% (for Praha, the most clustered) and 33% (for Liberec, the most uniform). The municipalities Praha, Brno, Olomouc tend to occurre as clustered. On the other hand, Liberec, Pardubice and Karlovy Vary figure more uniformly. The differences (except Praha) are small.

8 Conclusion

The paper presents a possible way how to process media news especially if a RSS channel exists. The first results of processing CT24 Regional News declare the roughly even distribution of means, nevertheless some deviations were recognised – i.e. Karlovy Vary seems to attract more news than other towns.
The longer period of monitoring and analysing will provide more significant results.
The analysis of time distribution did not bring any significant differences in time clustering for selected large cities.

Using methods mentioned above it is possible to reach higher accuracy of geocoding of the text information. It would require to implement more detailed locations (such as parts of towns, streets) to reference data. In addition to this, it is possible to extend reference data by connecting geonames: mountains, water courses or vegetation zones.
The challenge is to analyse and geocode audio, graphic or video information. In these analyses it is inevitable to use more sophisticated methods such as voice analysis.
An application of GeoRSS containing geographical coordinates seems to be a good choice for geocoding text information. The weakness is that GeoRSS is primary aimed for particular types of events, not media news. Furthermore, GeoRSS is not widespread in the Czech Republic.

References

1. BAILEY T.C., GATRELL A.C. *Interactive spatial data analysis*. Essex, Longman Scientific & Technical, 1995
2. Beaman, Reed S. and Conn, Barry J. *Geoparsing and georeferencing of Malesian collection locality data.* Telopea, 2003. 10(1) 43–52.
3. Bella, T. *Prvá slovenská mapa spravodajstva.* http://bella.blog.sme.sk/c/109353/Prva-slovenska-mapa-spravodajstva.html
4. CHARVÁT, Karel- Kocáb, Milan-Konečný, Milan- Kubíček, Petr. *Geografická data v informační společnosti.* VÚGTK, 2007, Prague. p. 42-44. ISBN 970-80-85881-28-8
5. HOLZNER, Steve- Šindelář, Jan. *RSS: Automatické doručování obsahu vašich WWW stránek.* Computer Press, 2007 Brno. ISBN 978-80-251-1479-7
6. HORÁK, Jiří. *Prostorová analýza dat.* Institute of geoinformatics, 2006, Ostrava. p. 37-38
7. Mikaela Keller, John S. Brownstein, Clark C. Freifeld 2008. Expanding a Gazetteer-Based Approach for Geo-Parsing Disease Alerts. (prior-knowledge-language-ws.wdfiles.com/local--files/start/keller_slides.pdf)
8. RSS Specifications: History of RSS. http://www.rss-specifications.com/history-rss.htm
9. The Berkman Center for Internet & Society at Harvard Law School. RSS 2.0 Specification. http://cyber.law.harvard.edu/rss/rss.html#requiredChannelElements