Combining GIS and R for dialectal data analysis and representation

Kristel Uiboaed
Maarja-Liisa Pilvik
Eleri Aedmaa
Siim Antso
University of Tartu

ARCHIMEDES

Outline of the presentation

- About the project “Applying Spatial Data in Linguistics”
- The aim of the presentation
- About the Corpus of Estonian Dialects (CED)
- Digitizing dialect atlas data
- Analyzing and visualizing data with R
- Using R to combine data from different sources
- Conclusions

About the project (RuRaKe)

- http://rurake.keeleressursid.ee/
- “Ruumiandmete rakendamine keeleteaduses” (Applying Spatial Data in Linguistics)
- Initially:
  - Adding geographical information to the Corpus of Estonian Dialects
  - Digitizing dialect maps of the “Väike Eesti murdeatlas” (“Small atlas of Estonian dialects”, 1955) by Andrus Saareste
- Creating resources for combining and analyzing different data sources with a spatial dimension
- Bringing dialect data closer to people not actively working in the field

The aim of the presentation

- Show, how combining different data sources (atlas data and corpus data) and applications (GIS-systems and statistical computing software R) is beneficial for getting a more diverse picture of linguistic variation and the language phenomena
- 3 alternation phenomena:
  - *sui* / *suvi* ‘summer’ (Saareste 1955)
  - *suurem* / *suuremb* ‘bigger’ (Saareste 1955)
  - *ei ole* / *pole* ‘is not’ (Saareste 1955)

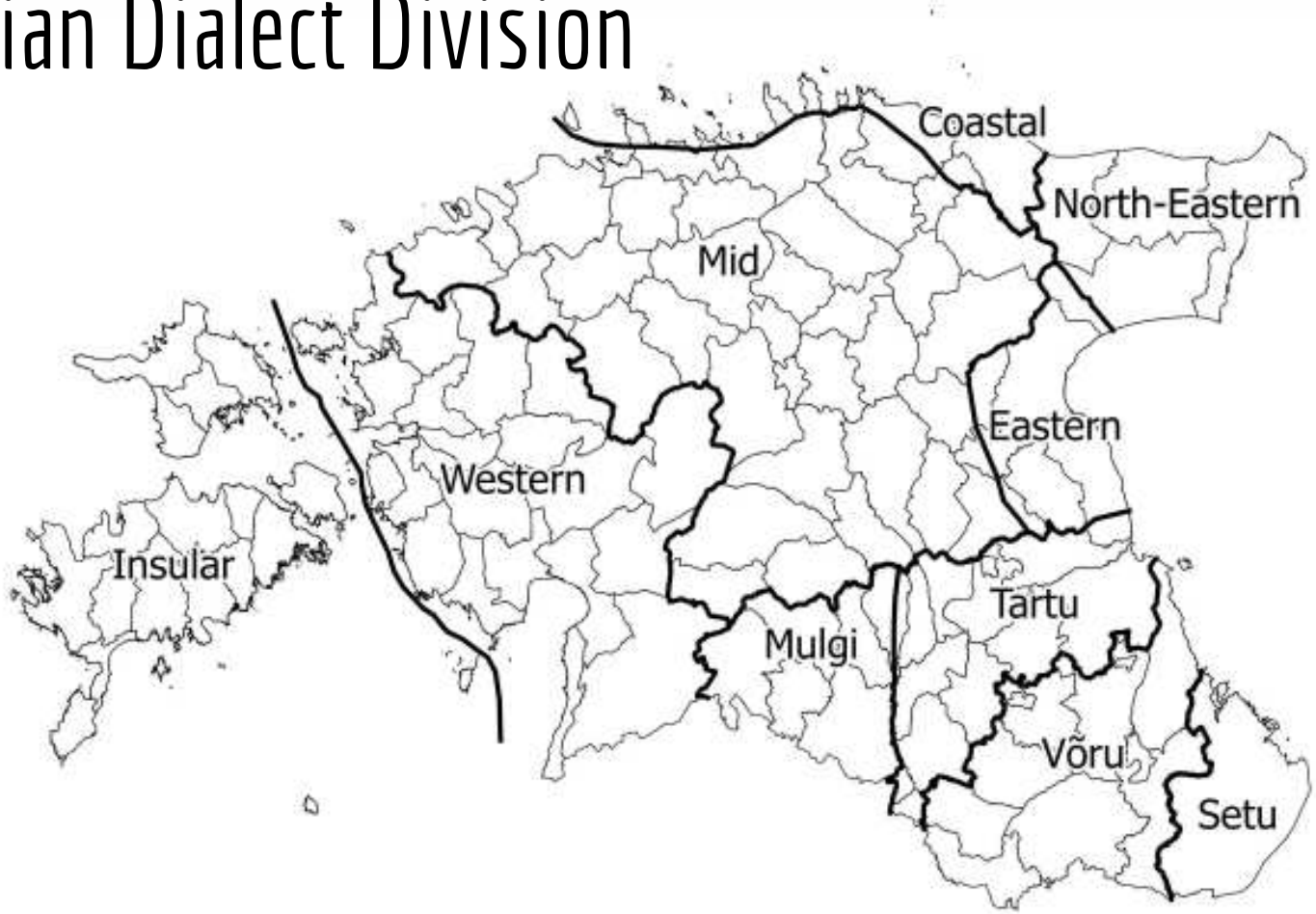
About the Corpus of Estonian Dialects

- http://www.murre.ut.ee/mkweb
- The electronic corpus contains text from all the Estonian dialects.
- The corpus mainly consists of transcribed and morphologically annotated texts of interview recordings from the 1960s and 1970s, as well as the recordings themselves and the metadata.
- Morphological annotation is done manually. Annotation includes information about lemma, part-of-speech, morphological categories, and meanings.

HLJ_Mare_Pihlakas_EMH0135_synt(UTF-8).xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="liivike_tekst.xsl"?>
<liivike>
    <info><tyyp>suuline</tyyp><kihelkond>Hlj</kihelkond><keel>eesti</keel><murre>Ranna</murre><kirjaviis>lihtsustatud transkriptsioon</kirjaviis><muu>Rannikumurre, Haljala murrak, Potsu küla, Marie Pihlakas (85-aastane). Lindistanud (EMH 135, 1959. aastal) ja litereerinud (KKI MT 238) Mari Must (MM), üle kuulanud ja täiendanud M-L. Kalvik 2008. aastal. KJ ? Marie Pihlakas, ? ? meesterahvas, kes juttu eenalt täiendab.</muu><lindistus lindistaja="Mari Must" aasta="1959">EMH0135</lindistus><litereering litereerija="Mari Must" aasta="1959">KKI MT 238</litereering><kyla longituud="26.20667" latituud="59.46056">Potsu</kyla><intervjueerija id="?" haridus="NA" synniaasta="NA" synnikoht="NA" vanus="NA" sugu="M">meesterahvas, kes juttu eenalt täiendab</intervjueerija><keelejuht id="KJ" haridus="NA" synniaasta="1874" synnikoht="Potsu" vanus="85" sugu="N">Marie Pihlakas</keelejuht><intervjueerija id="MM" haridus="kõrg" synniaasta="1920" synnikoht="Tartu" vanus="39" sugu="N">Mari Must</intervjueerija></info><sisu id="HLJ_Marie_Pihlakas_EMH0135_synt">
        <lause id="l1" koneleja="KJ"><sone id="l1_s1" meta="vahemärk">(=)</sone><sone id="l1_s2" lemma="olema" vorm="pers.ind.ipf.sg.3." liik="V">oli</sone></lause><lause id="l2" koneleja="MM"><sone id="l2_s1" meta="intervjueerija">jahh</sone><sone id="l2_s2" meta="vahemärk">(.)</sone></lause><lause id="l3" koneleja="KJ"><sone id="l3_s1" lemma="ja" liik="Konj">ja</sone><sone id="l3_s2" lemma="siis" liik="ProAdv">siis</sone><sone id="l3_s3" meta="vahemärk">=</sone><sone id="l3_s4" lemma="e" liik="Par">e</sone><sone id="l3_s5" meta="vahemärk">(...)</sone><sone id="l3_s6" lemma="utlema" vorm="pers.ind.ipf.sg.3." liik="V">*utles</sone><sone id="l3_s7" lemma="veel" liik="Adv">viel</sone><sone id="l3_s8" lemma="ena" vorm="sg.all." liik="S">enale</sone><sone id="l3_s9" meta="vahemärk">(.)</sone><sone id="l3_s10" lemma="enne" liik="Adv">*enne</sone><sone id="l3_s11" lemma="et" liik="Konj">et</sone><sone id="l3_s12" meta="vahemärk">=</sone><sone id="l3_s13" lemma="e" liik="Par">e</sone><sone id="l3_s14" meta="vahemärk">(...)</sone><sone id="l3_s15" lemma="ära" vorm="pers.imp.pr.sg2." liik="Mn">ära</sone><sone id="l3_s16" lemma="siis" liik="Par">siis</sone><sone id="l3_s17" lemma="laps" vorm="pl.part." liik="S">*lapsi</sone><sone id="l3_s18" lemma="laskma" vorm="pers.imp.pr.neg." liik="V">lase</sone><sone id="l3_s19" lemma="hunt" vorm="sg.el." liik="S">*un'dist</sone><sone id="l3_s20" lemma="minema" vorm="inf." liik="V">*menna</sone><sone id="l3_s21" lemma="et" liik="Konj">et</sone><sone id="l3_s22" lemma="karistama" vorm="pers.imp.pr.sg.2." liik="V">karista</sone><sone id="l3_s23" lemma="nemad" vorm="pl.part." liik="ProS">neid</sone><sone id="l3_s24" lemma="ka" liik="Par">kaa</sone><sone id="l3_s25" meta="vahemärk">(.)</sone></lause>
<lause id="l4" koneleja="MM"><sone id="l4_s1" meta="intervjueerija">nii</sone></lause>
```

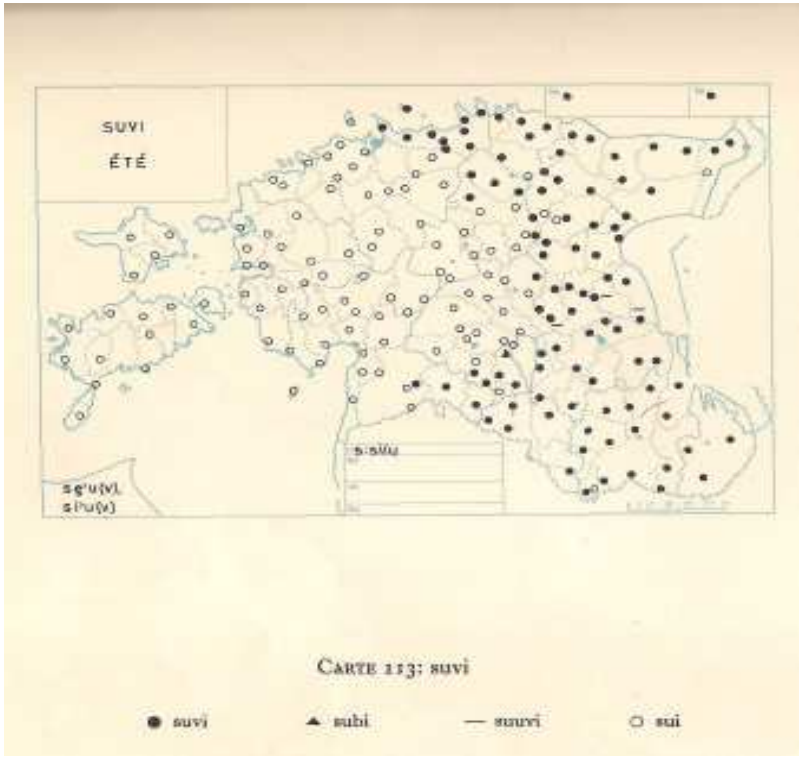
Estonian Dialect Division



Atlas data and using GIS

- A. Saareste's two atlases:
  - 1938/1941 "Eesti murdeatlas" ("Atlas of Estonian Dialects") with 66 maps
  - 1955 "Väike Eesti murdeatlas" ("Small Atlas of Estonian Dialects") with 125 maps
- "Atlas of Estonian Dialects" (AED) contains maps with text (a la Jules Gilliéron)
- In "Small Atlas of Estonian Dialects" (SAED) Saareste also used map symbols to visualize dialect data
- Mostly maps of vocabulary in AED (about 70%)
- SAED contains more morphology and phonology (57%) as well as vocabulary (41%)
- Using QGIS
  - Scanning and georeferencing
  - Assigning phonetically transcribed linguistic data (words, phrases, constructs) to the centroids of villages where the data was collected

Andrus Saareste's dialect maps



CARTE 30: ei ole < e(p)i ole-k

| ei (ãi) ole	● pole	× ei (ãi, õi) pole	▽ -p põle
— ei oole	○ põle	+ ei põle	⊥ ole ei (õi)
∧ ei õle	△ põle	▲ -p ole (õle)	⊤ ei olõ' õijõ'
S ep ole	◐ põlõ	▼ -p pole	⌒ oleki
			Y ãle

Using R

Perks of using R for dialectal data manipulation:

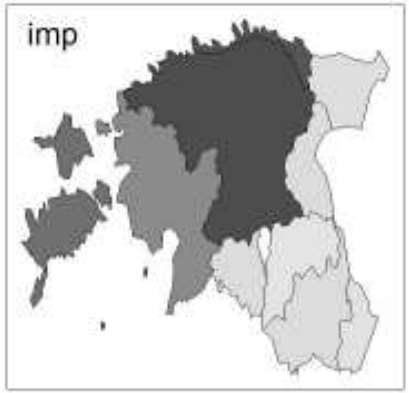
- Ready made analysis tools support limited number of statistics
- Possibility to manipulate dialect borders
- All the steps of the analysis are in the same environment
- A large number of statistical methods is supported in R (9144 packages available at http://cran.r-project.org/web/packages/available_packages_by_date.html)

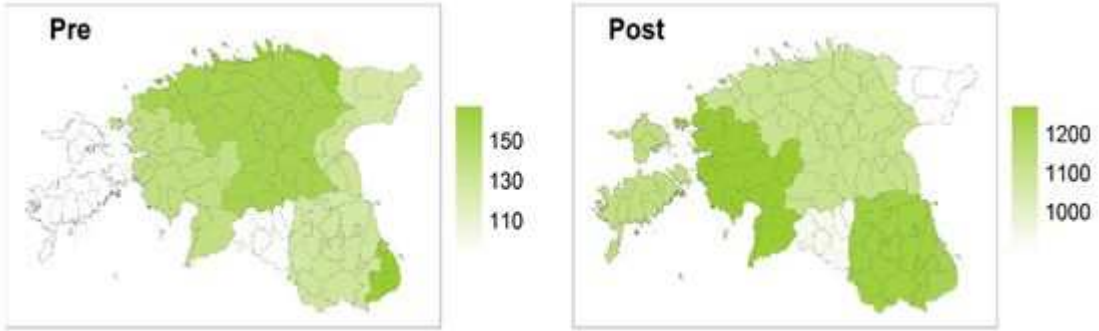
R packages

Packages for visualization:
Ggplot2 (Wickham 2009)
Ggmap (Kahle & Wickham 2013)

Packages for processing geodata:
Rgdal (Bivand et al. 2015)
Rgeos (Bivand & Rundel 2015)

Studies based on CED data

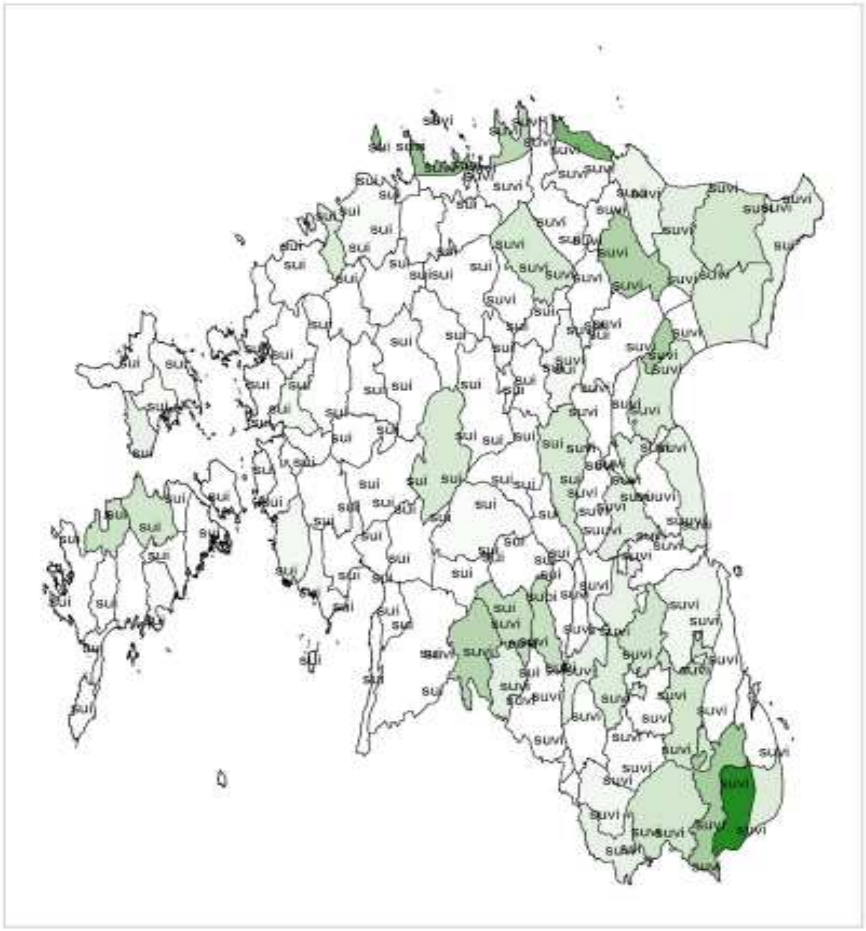


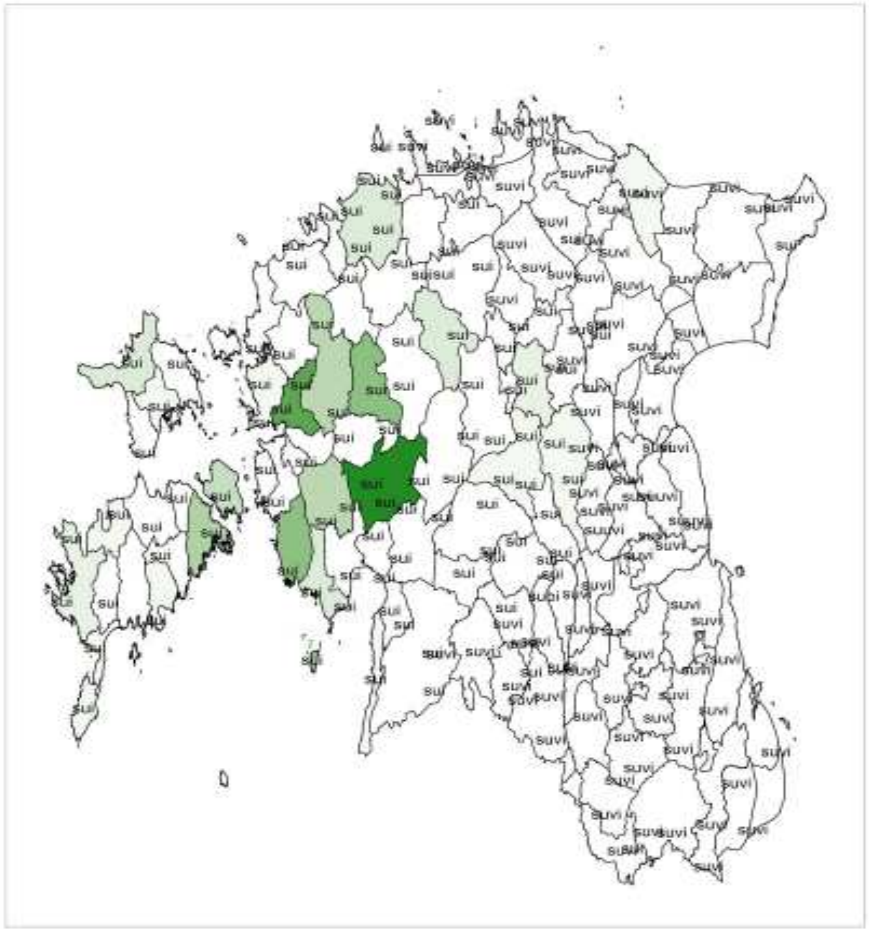


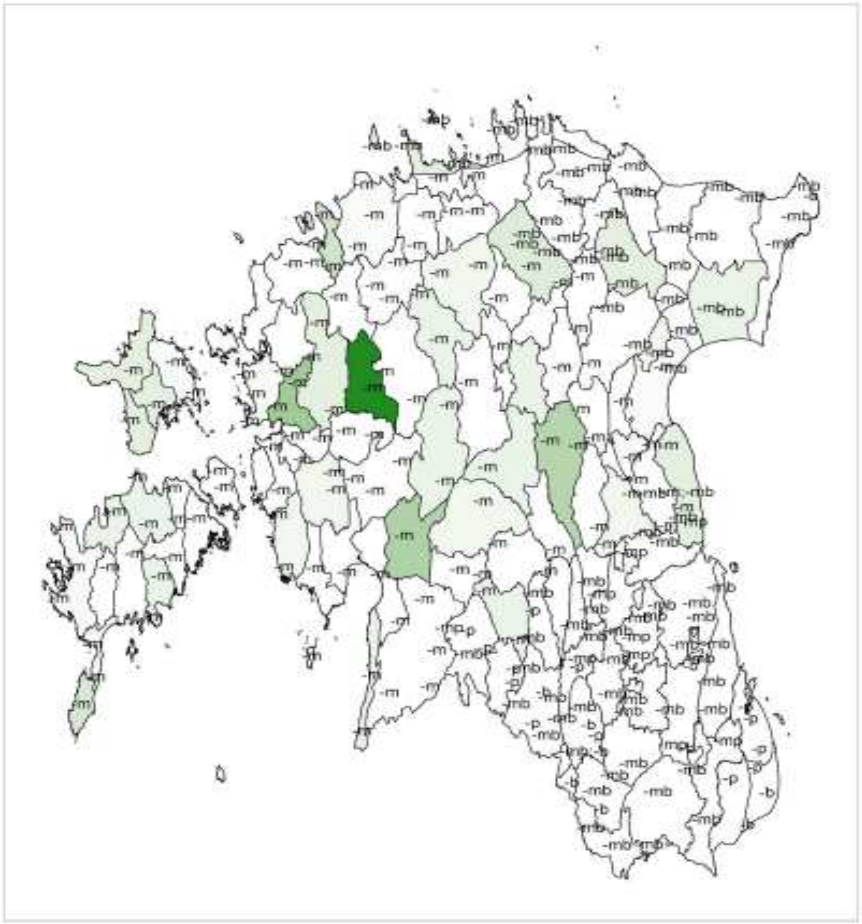
- Uiboaed, Kristel (2013): *Verbiühendid eesti murretes.* Tartu: Tartu Ülikooli Kirjastus. (= Dissertationes Philologiae Estonicae Universitatis Tartuensis).
- Vihman, Virve-Anneli, Liina Lindström, and Kristel Uiboaed (2014): "Varieerumine tarvis-/vaja-konstruktsioonides keelekontaktide valguses." *Keel ja Kirjandus* 8-9, pp. 609-630.
- Ruutma, Mirjam. *Kaassõnad eesti murretes*. Diss. Tartu Ülikool, 2016.
- Uiboaed, Kristel 2016. *Spatial Visualization with R.* spatial-visualization-with-r 1.0. http://dx.doi.org/10.5281/zenodo.51473.

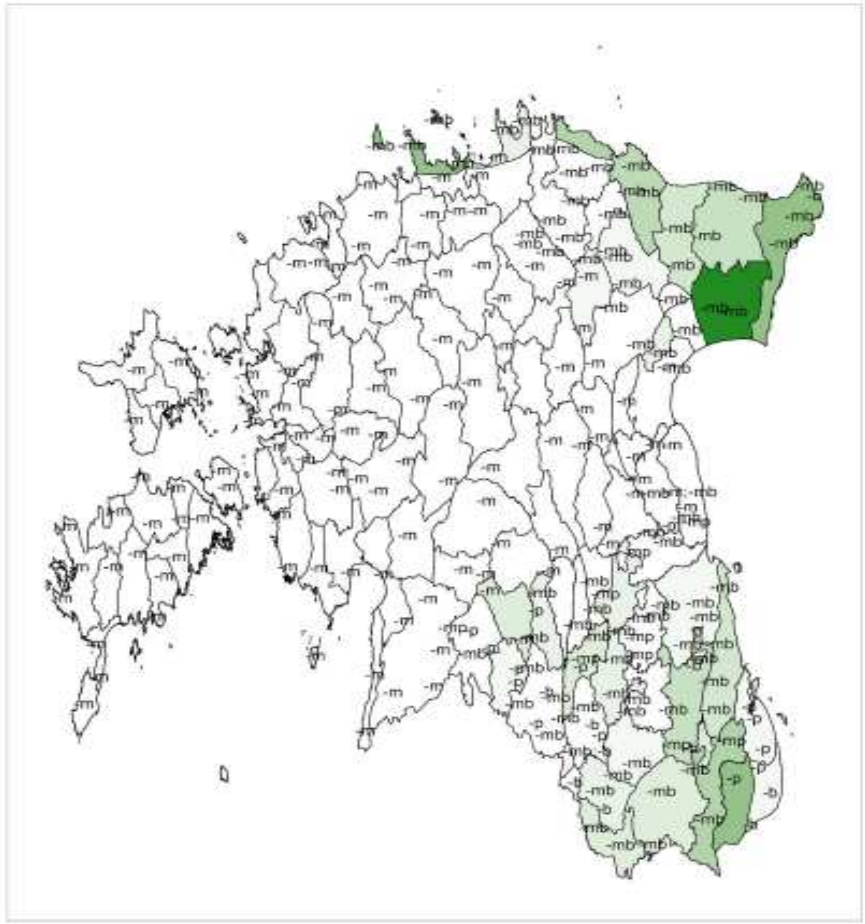
Corpus data

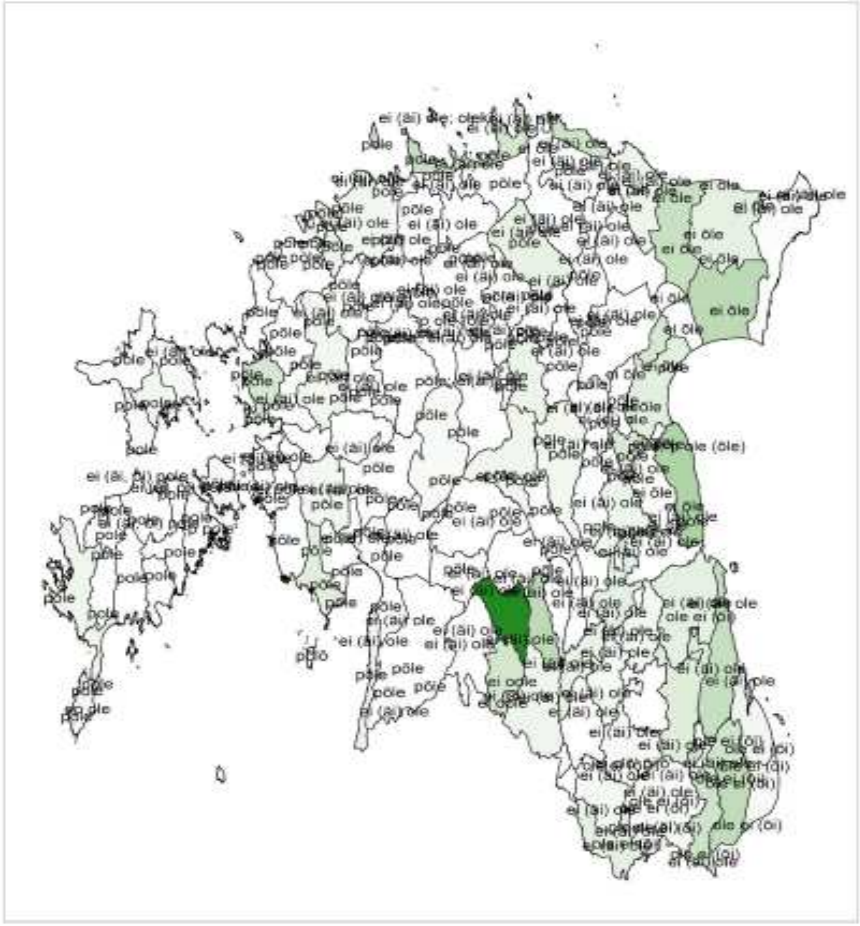
- 973 147 morphologically annotated running words
- *suvi / sui*
  - 375 occurrences from 62 parishes
  - Frequencies normalized based on a mean corpus size
- *suurem / suuremb*
  - 218 occurrences from 62 parishes
  - Frequencies normalized based on a mean corpus size
- *ei ole / pole*
  - 2969 occurrences from 53 parishes
  - Data drawn from a balanced sample of ~40 000 words/dialect (total of 405 287 words)

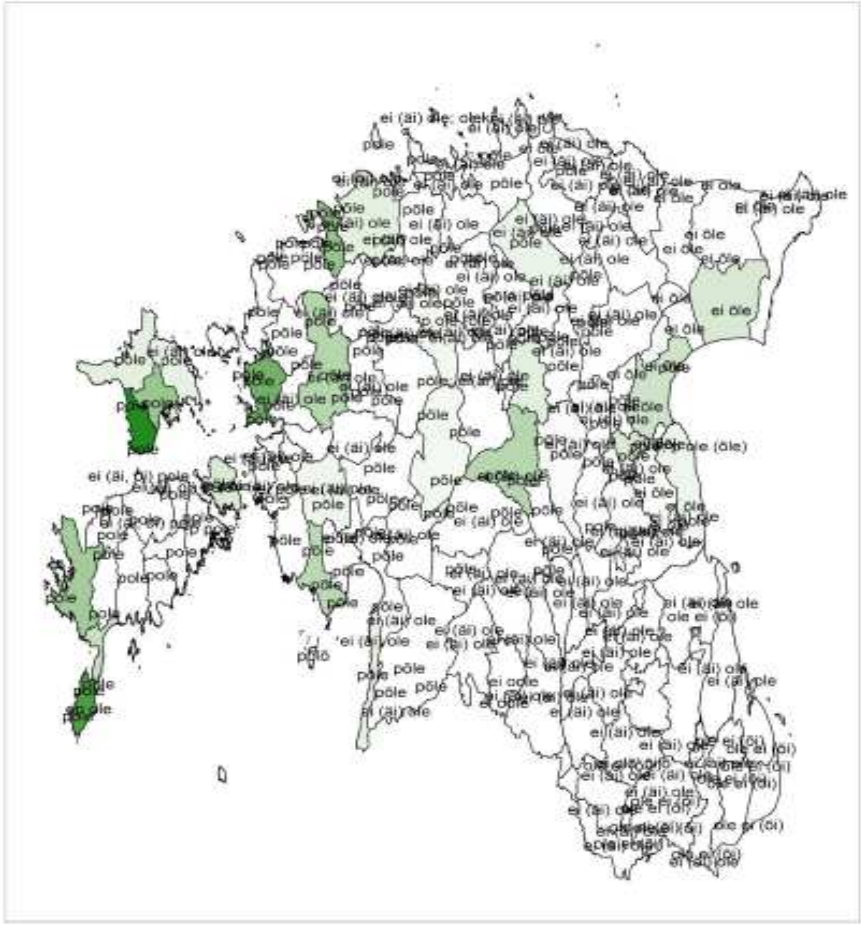

suvi
sui


suvi
sui






pôle
ei (ai) ole
ei ôle
ei (ai, ôi) ole
ei (ai) ole (ôle)


pôle
ei (ai) ole
ei ôle
ei (ai, ôi) ole
ei (ai) ole (ôle)

Shiny

- An RStudio package for creating interactive web applications
- Relatively simple syntax
- No web programming knowledge required
- Enable everything that is possible also in R, among other things, interactive mapping and statistical analysis of dialectal data
- Visualizing dialect data for the “lay person”
- Initial prototypes available at http://rurake.keeleressursid.ee/index.php/apps/

In conclusion

- Dialectal data analysis has always had an inherent spatial component
- Corpus data provides frequency information, linguistic atlases information on the spread of certain linguistic phenomena
- Combining the 2 data sources provides us with a more diverse understanding of dialectal variation
- R enables to perform a wide range of operations, starting from data extraction all the way to geographic visualization
- In addition to making dialect data more accessible, Shiny applications would also be a suitable platform for crowdsourcing