Non-traditional data sources in Social Statistics of Statistics Finland

Pasi Piela, pasi.piela@stat.fi

Non-traditional data sources in the National Statistical Systems, 17th Meeting of ECLAC, Santiago de Chile, 1-2 October 2018

Contents

- Accessibility statistics
- Mobile network data
- Web-scraping
- Managerial view

Accessibility as a concept

- Still very relevant part of today’s geographic information science.
- This presentation does not include accessibility estimation for persons with disabilities.
- The UN Sustainable Development Goals are motivating towards such research at Statistics Finland too – together with other national stake holders. E.g.:
    - SDG 11.2.1: Proportion of population that has convenient access to public transport, by sex, age and persons with disabilities

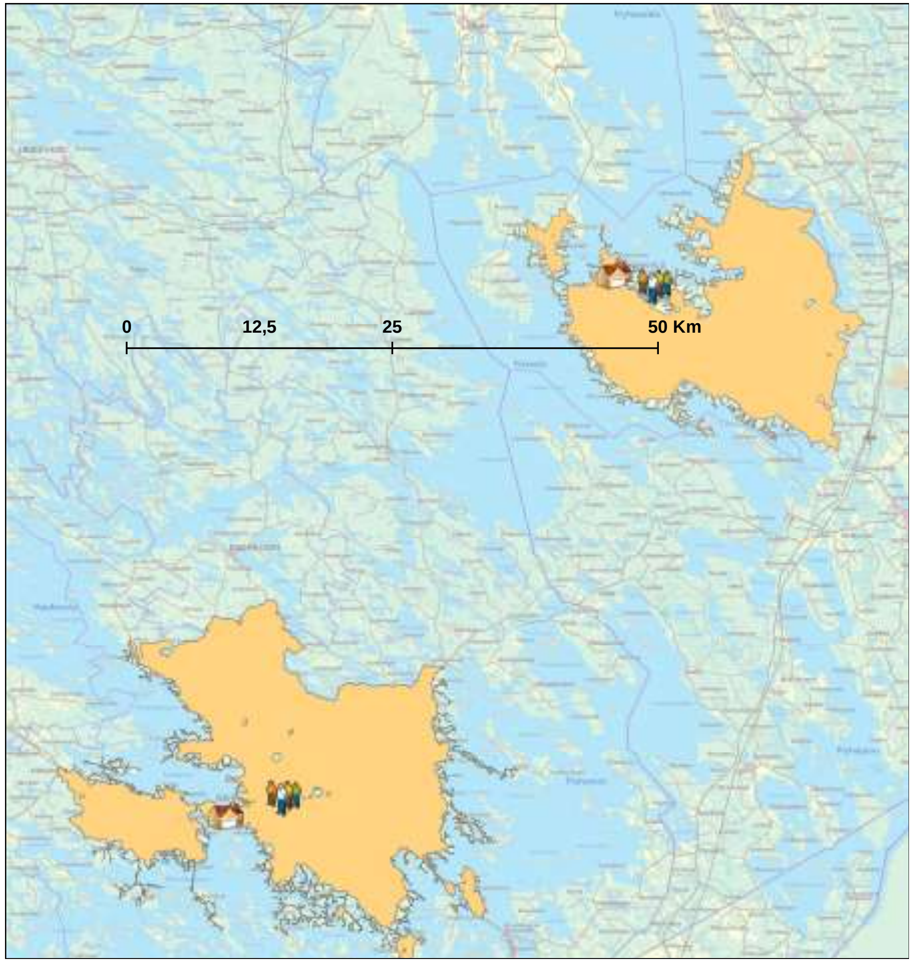
Spatial data sources of Social Statistics

- Plenty of administrative and register-based data available for many kinds of research on the population itself and of services it is potentially using.
- Combined to statistical products for customers of StatFi
- Special enquiries require data from customers: e.g. festivals in Finland
- Basic services: travel time and distance estimation from point to point by applying the Finnish National Road and Street Database *Digiroad (digiroad.fi)*.

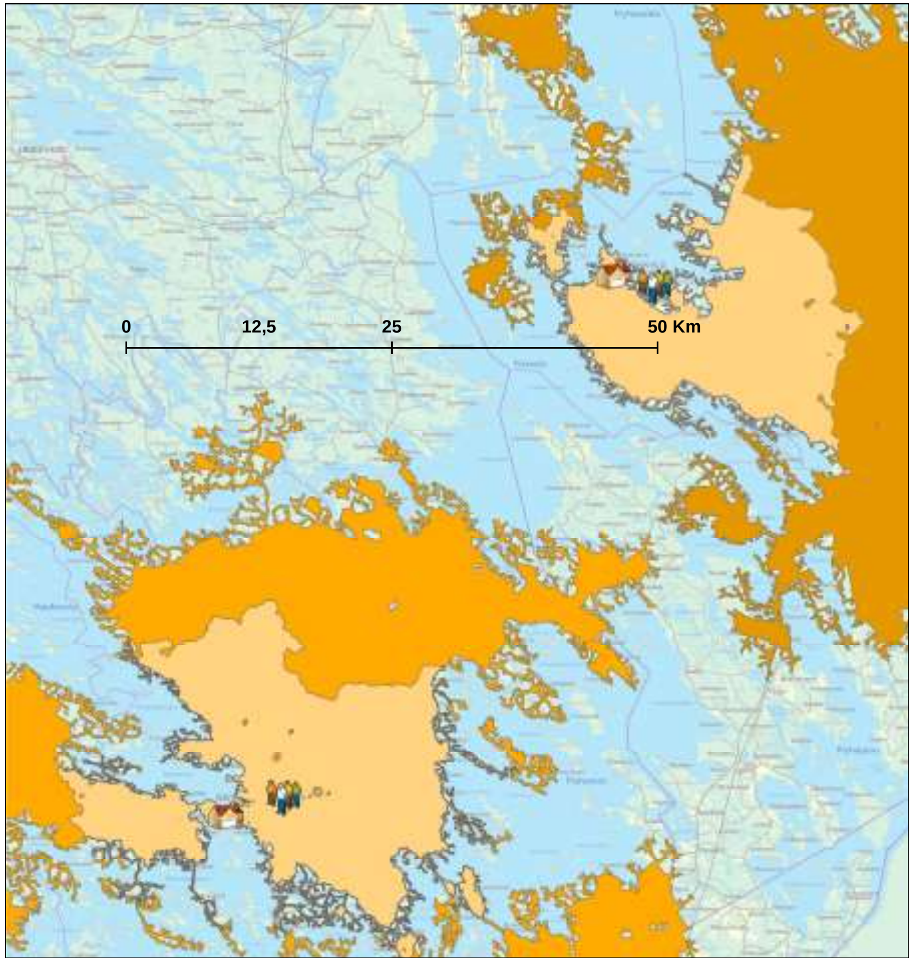
Remoteness (index) estimation, Ministry of Finance

- Part of the state subsidies to municipalities
- Currently a simplified system putting together 25 km and 50 km buffers around municipal population center points (by 1 km x 1 km population grids)
- Enrichment proposal: service area polygons around the municipal population center points (”trimming” 100 meters along roads, applying 250 m x 250 m population grids)

Savonlinna and Rääkkylä 25 km service area polygons around the population center points

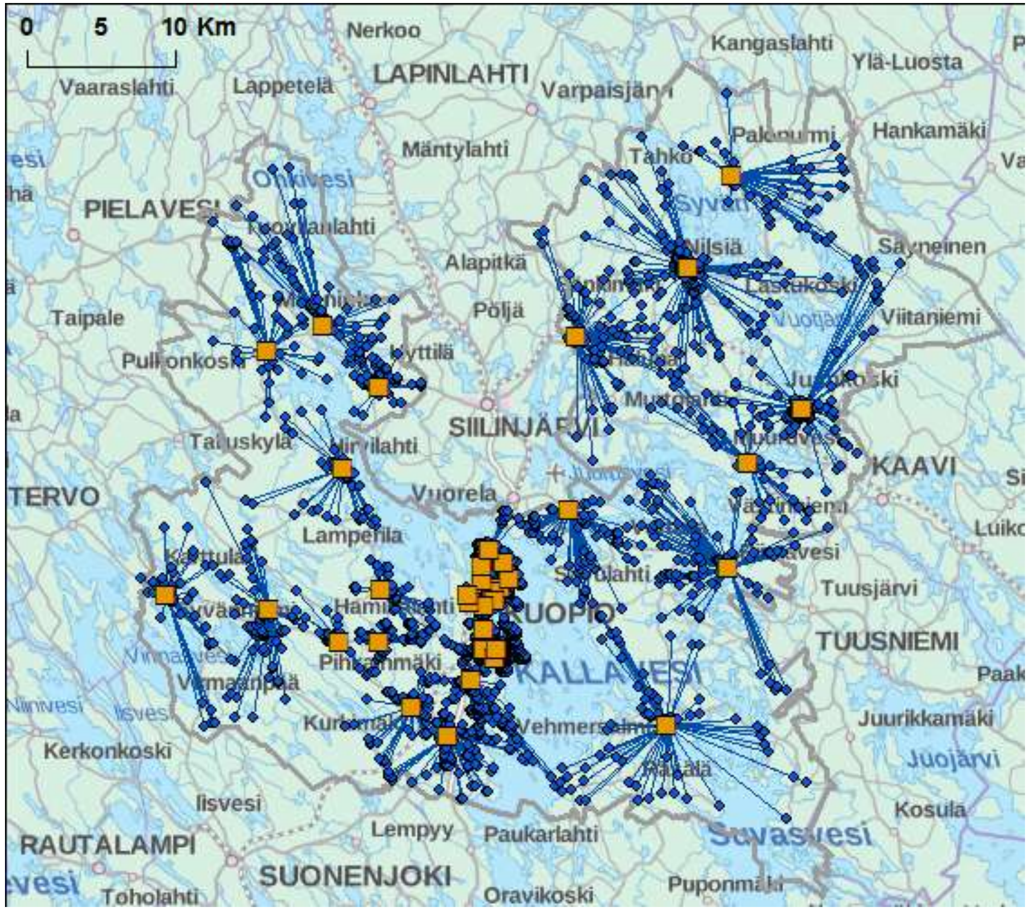


Savonlinna and Rääkkylä 25 and 50 km service area polygons around the population center points



Elementary school accessibility

- Annual, “simple”, point-to-point road distance estimation among school children (age groups separately)
- Private schooling irrelevant here



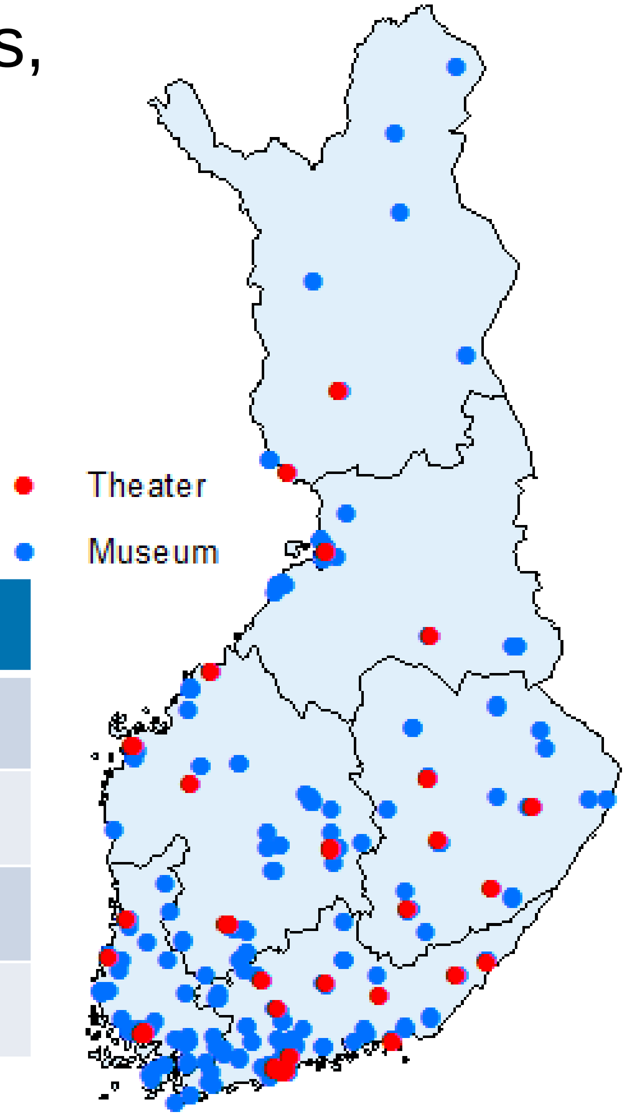
Cultural accessibility

- Many applications: libraries, theatres, movie theatres, orchestras, festivals, childrens' cultural centres etc.
  - Part of the cultural service data are collected by customers themselves
- Challenge: geocoding

Relative cultural accessibility in Finland:

	3 km	10 km	30 km
Festivals *	-	0.597	0.820
Theatres	0.200	0.500	0.715
Museums	0.331	0.679	0.881
Libraries	**0.724**	0.925	-

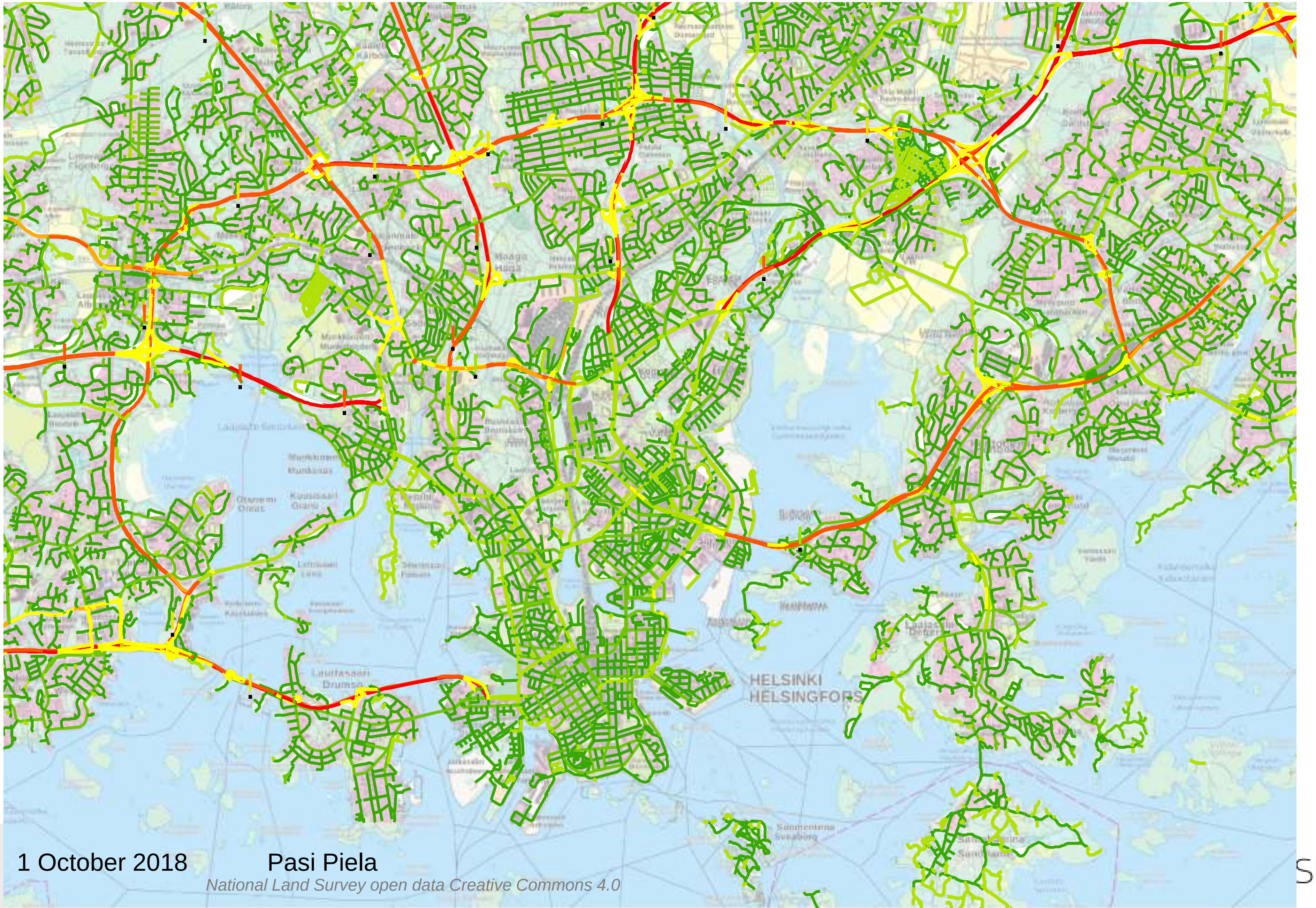
*) Finland Festivals & Statistics Finland



Commuting time estimation

- Data integration is based on many data sources, partly big data, in order to enrich official statistics of Finland. These include:
  - public transport data from web service platforms (APIs)
  - traffic sensor data
  - Digiroad
  - Plenty of administrative data
- National population coverage for the point-to-point estimation is about 93 %

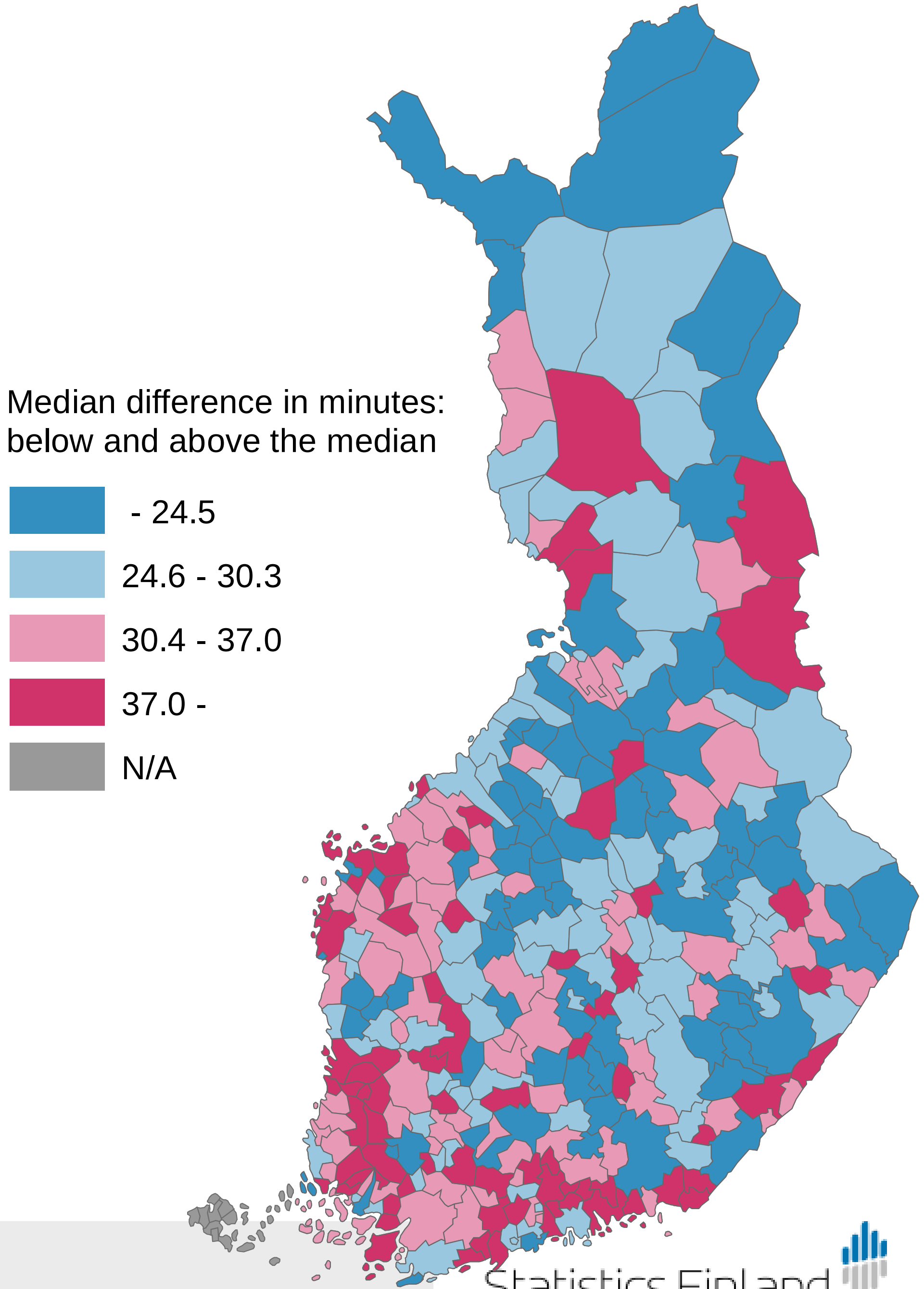
Automatic traffic measurement devices and speed estimates in Helsinki



National Land Survey open data Creative Commons 4.0

Commuting time estimation

- Municipal median differences of commuting times between the use of public transport and private car use:



Commuting time estimation

The new commuting database:

- Commuting distance and time by private vehicle,
- Cycling distance and time,
- Public transport distance and time,
- Helsinki Region Public Transport distance and time,
- Corrected commuting time for trips to and from the central Helsinki area.

Mobile network data

Mobile network data

- The leading example on big data in official statistics
- The most challenging e.g. due to **legal obstacles**
- Motivation in Finland comes from European examples and the work done within the European Statistical System community
- ESSNet Big Data project 2016-2018
  - https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data

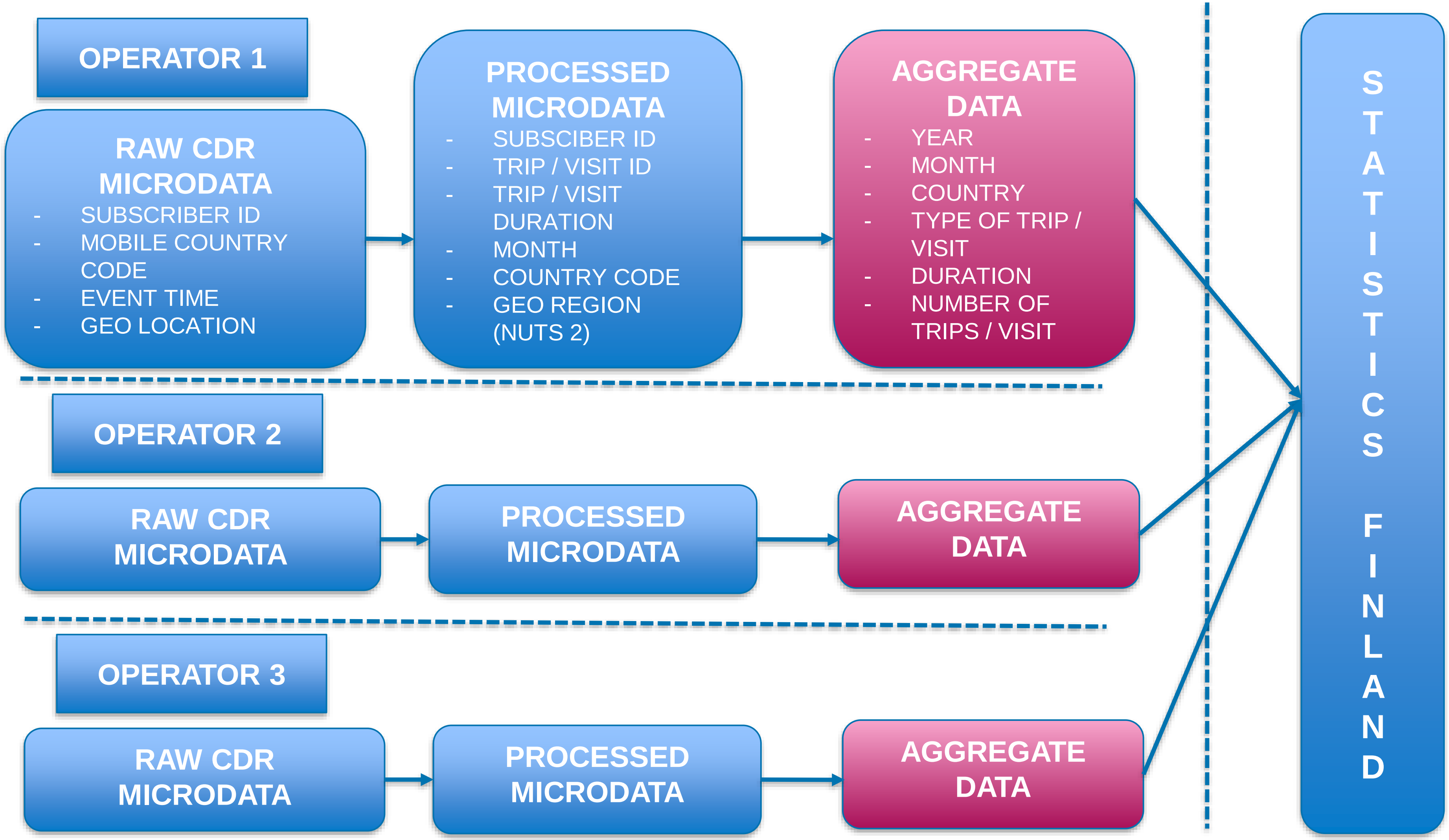
Mobile network data

- Priority is given to tourism statistics due to specific needs
- Seasonal population was secondary in this project, but it is needed, as not much information around on that topic except “Summer cottage statistics” – register/admin data collection
- Tourism statistics are presented here even though not part of the social statistics

Mobile data pilot for tourism statistics and for seasonal population

- Objective was to obtain pilot data from all three Finnish mobile network operators.
- a process description which details how aggregate tourism statistics can be compiled based on MNO CDR data
- covers inbound and outbound tourism; domestic tourism is currently out of scope
- Seasonal population covers the population estimation during certain weekdays and weekends on January and during the main summer holiday season (on July).
- Pilot has made progress with 2 out of 3 Finnish MNOs.

Process description



OPERATOR 1

RAW CDR MICRODATA
- SUBSCRIBER ID
- MOBILE COUNTRY CODE
- EVENT TIME
- GEO LOCATION

PROCESSED MICRODATA
- SUBSCRIBER ID
- TRIP / VISIT ID
- TRIP / VISIT DURATION
- MONTH
- COUNTRY CODE
- GEO REGION (NUTS 2)

AGGREGATE DATA
- YEAR
- MONTH
- COUNTRY
- TYPE OF TRIP / VISIT
- DURATION
- NUMBER OF TRIPS / VISIT

OPERATOR 2

RAW CDR MICRODATA

PROCESSED MICRODATA

AGGREGATE DATA

OPERATOR 3

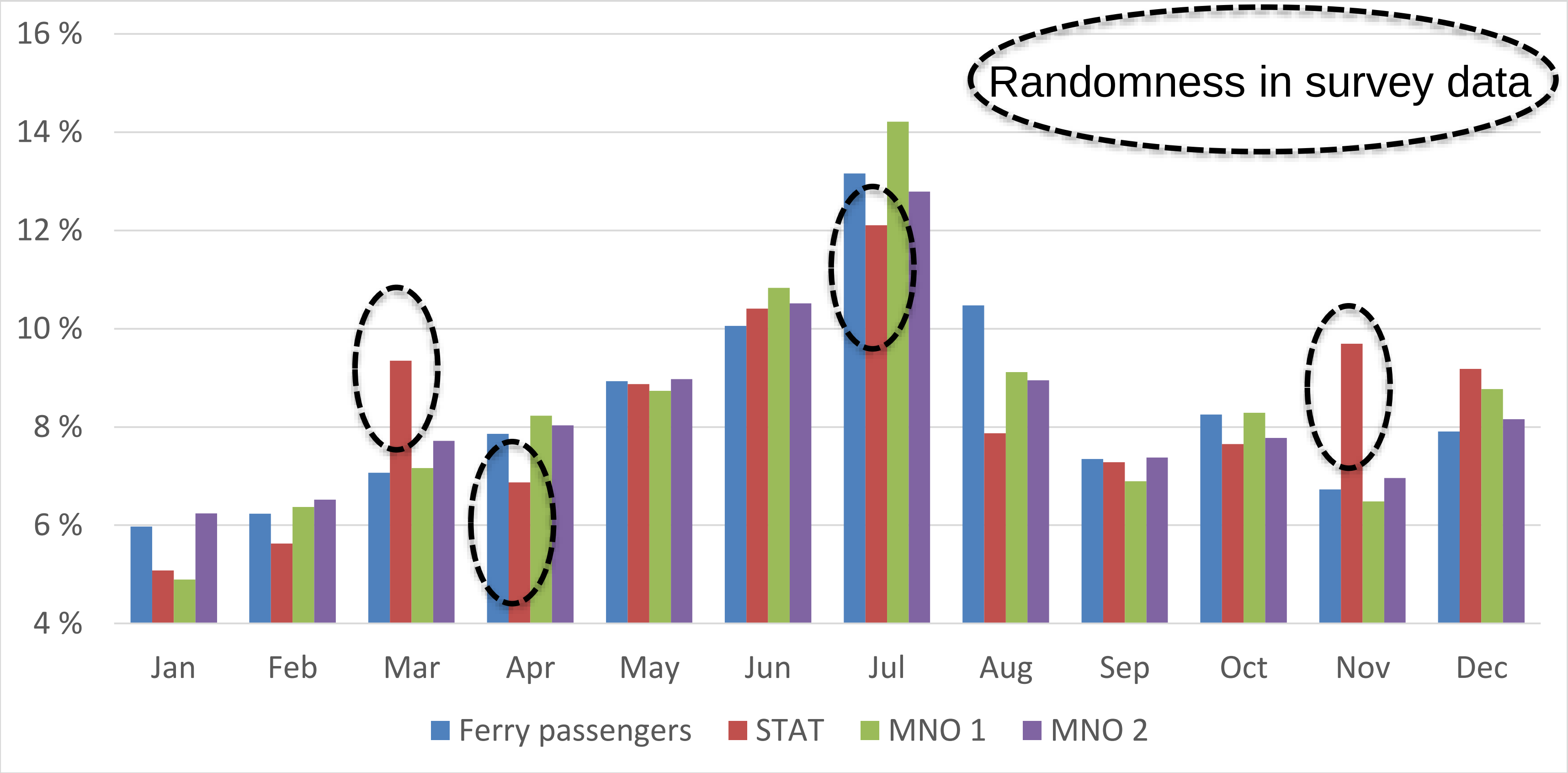
RAW CDR MICRODATA

PROCESSED MICRODATA

AGGREGATE DATA

STATISTICS FINLAND

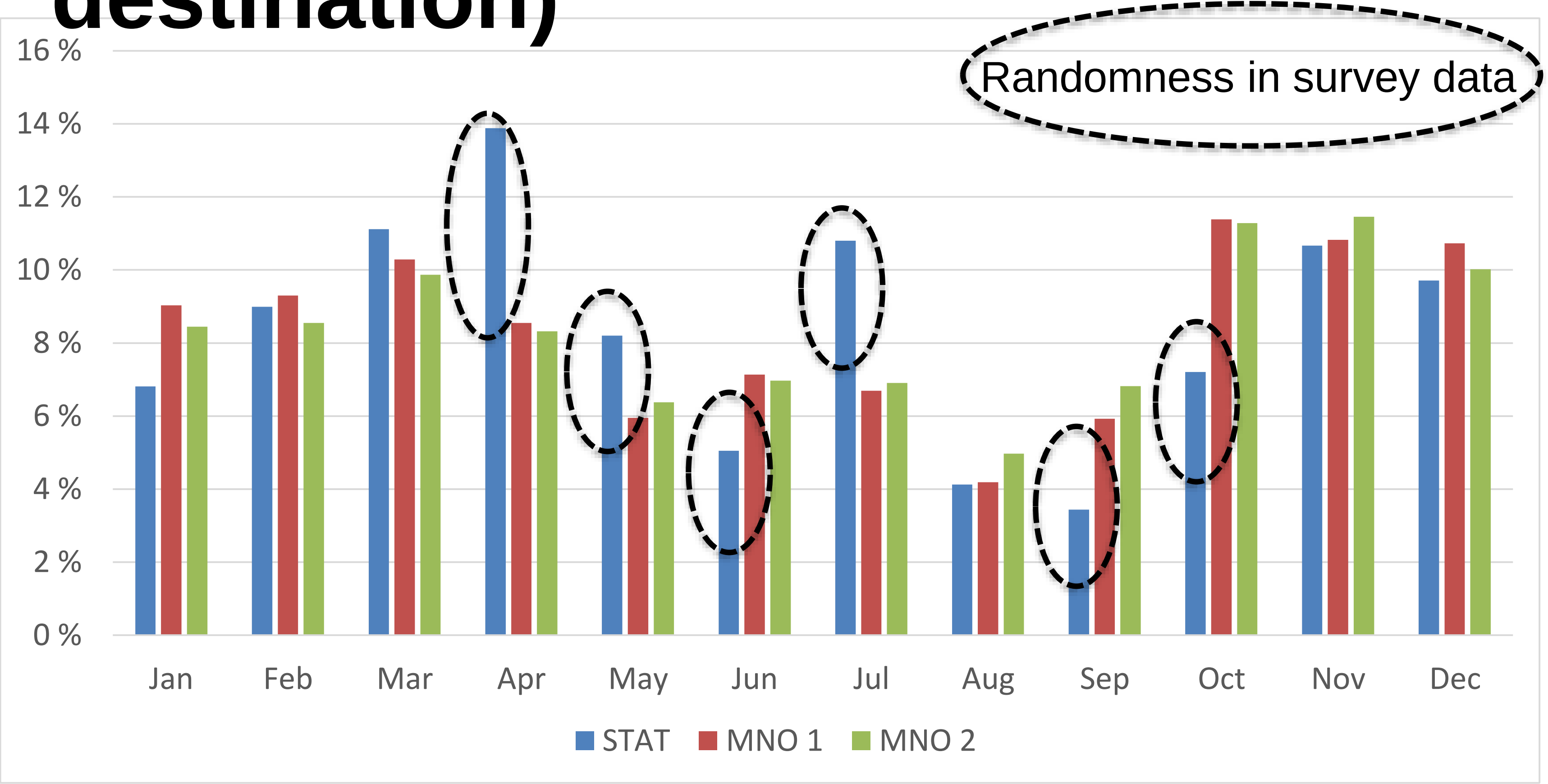
Outbound trips to Estonia



Helsinki is now the busiest passenger port of the world with 12 million people.

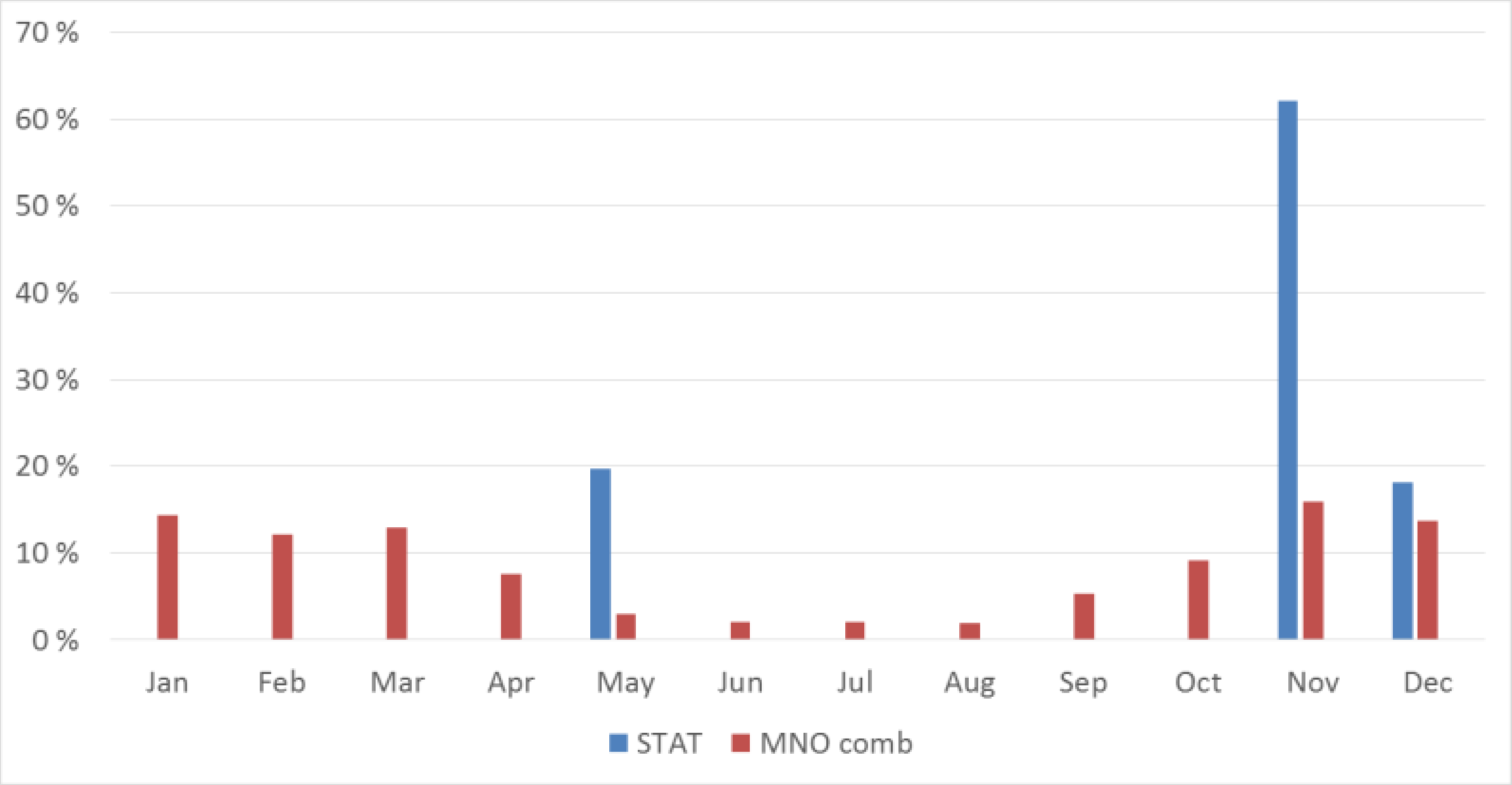
All data soures are mostly in consensus, but survey data is affected by randomness -> estimate is often too much or too little

Outbound trips to Spain (Top 3 destination)



MNOs are in consensus with each other, they differ only 0,5% units.
Survey trips are greatly affected by randomness.

Outbound trips to Chile

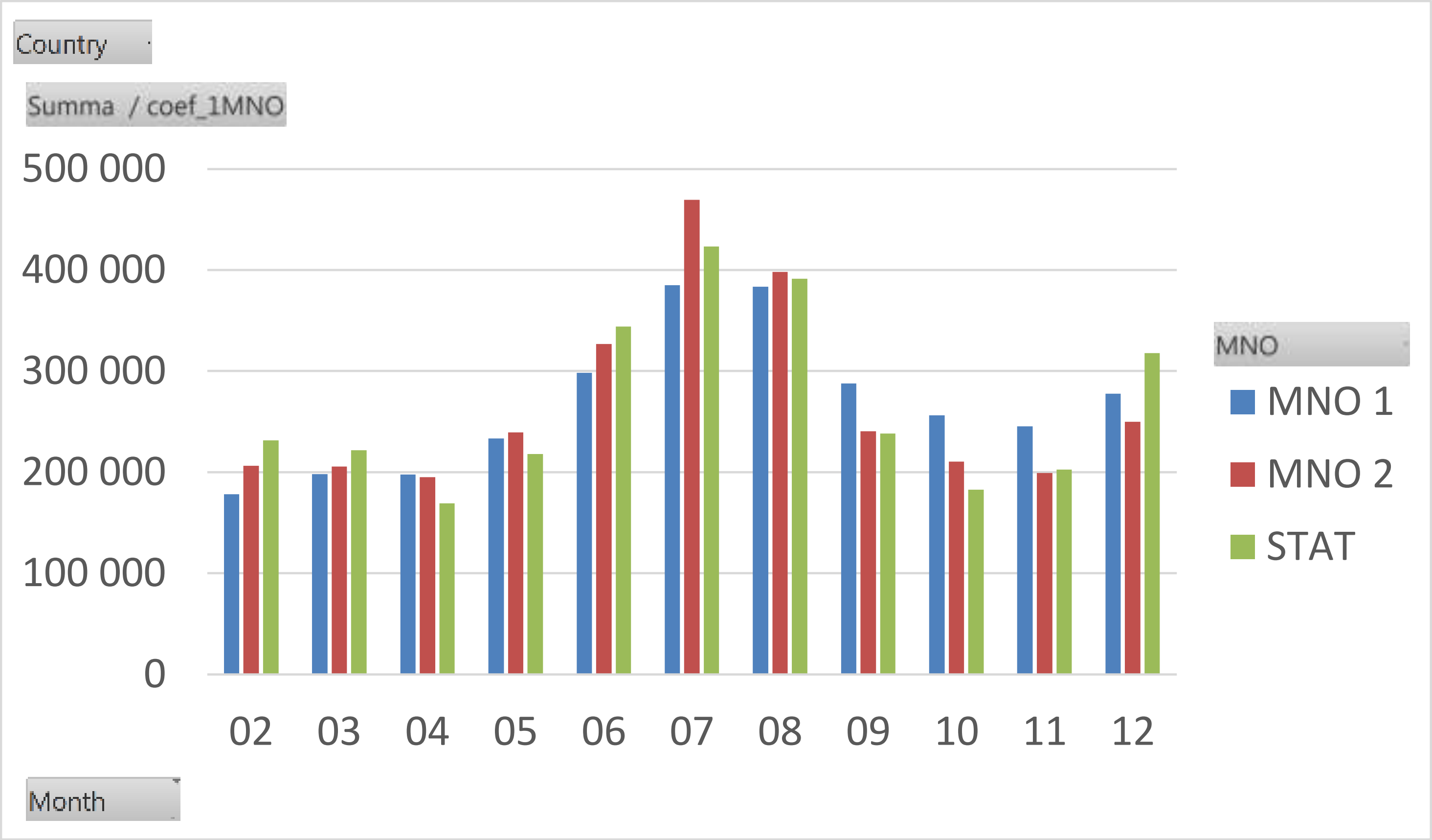


MNOs combined.

Outbound tourism conclusions

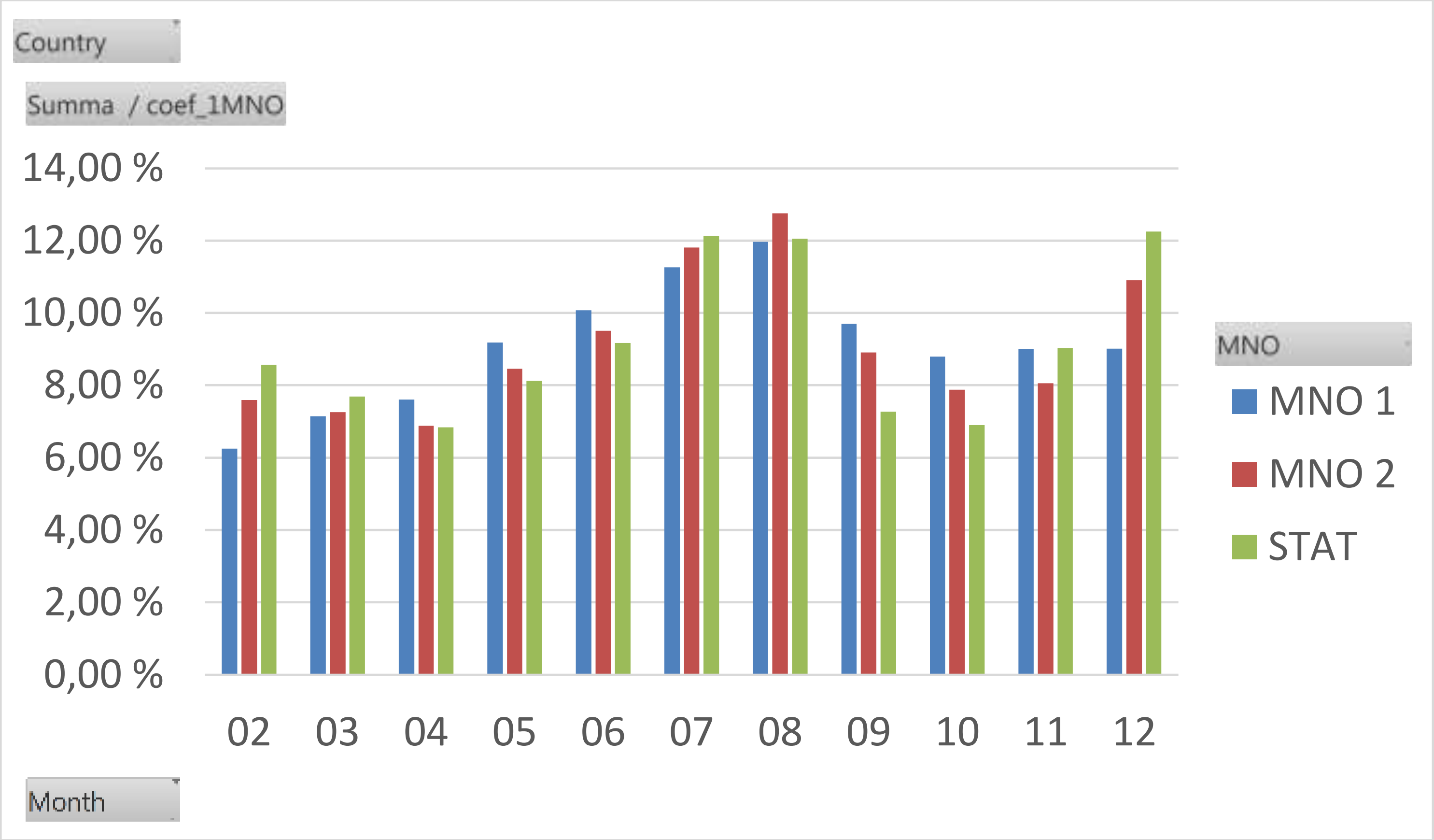
- The two MNOs have independently of each other provided data for outbound tourism
- MNO outbound data sets are in consensus with each other
- MNO data sets are describing the same ’elephant’
- There is high correlation to survey data also…
- …but survey is affected by randomness
- Smaller the destination -> less trips -> more randomness
- Preliminary conclusion – MNO outbound data should be used to mitigate randomness in the survey data

Monthly inbound tourism 2017

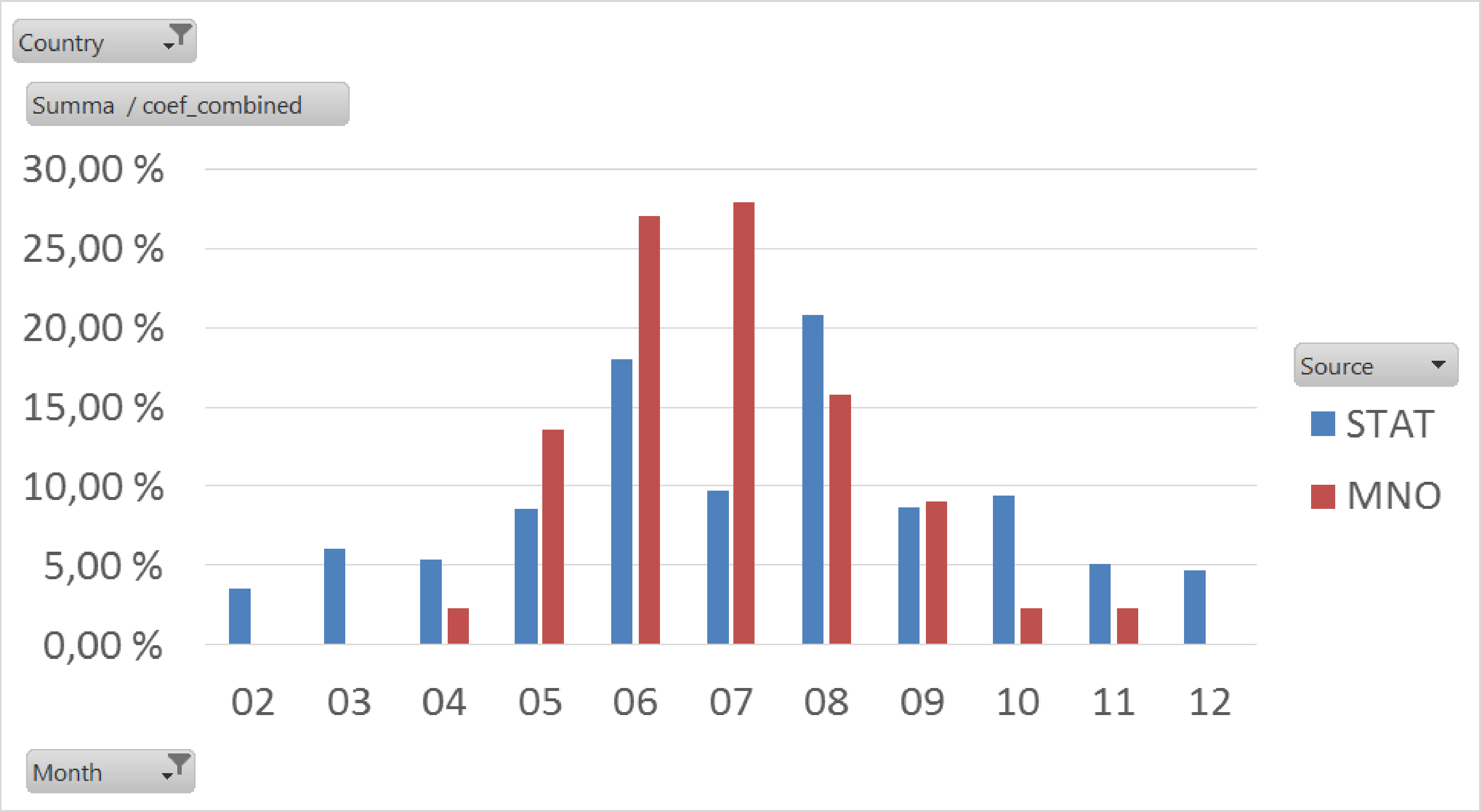


There is general consensus on inbound tourism monthly season in all sources.

Inbound trips from Russia



Inbound trips from Chile



MNOs combined.

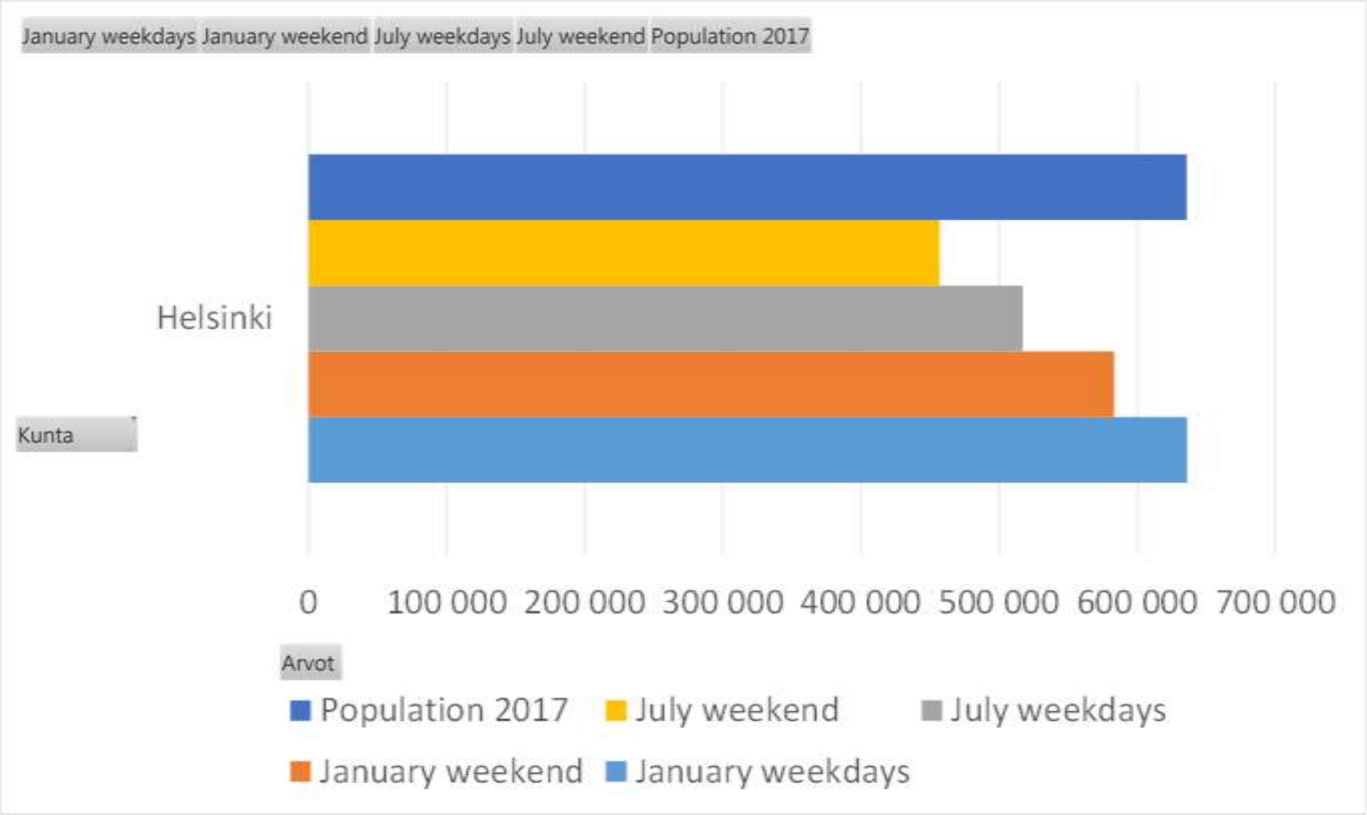
Inbound tourism conclusions

- There is a general consensus on monthly seasonality
- MNOs have different market shares depending on country of origin -> data from all 3 MNOs is needed for full picture
- Neighboring countries (EE, SE, NO, RU) have far more trips in MNO data than in accommodation statistics.
- Main inbound countries Japan and China seem to be underrepresented in MNO data?

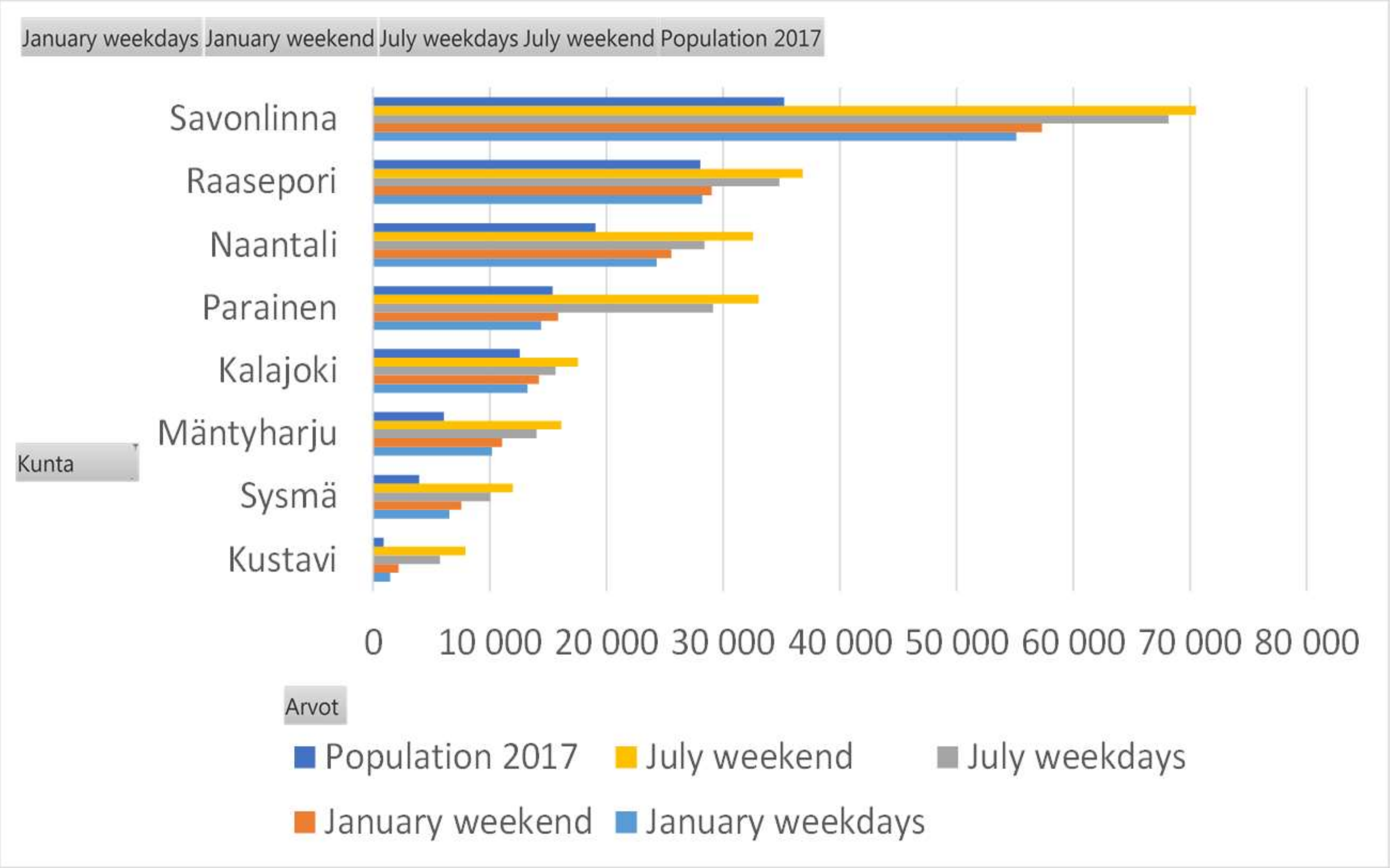
Mobile data for estimating seasonal population

- Mobile positioning data for seasonal population contains number of subscribers by municipality in Finland
- Data has been provided by two Finnish mobile network operators
- There are four different time periods
- Weekdays in winter (January)
- Weekend in winter (January)
- Weekdays in summer (July)
- Weekend in summer (July)
- Each subscriber is assigned to the municipality with the greatest number of transactions (call / sms / data) within the period
- Data from operators have been combined and extrapolated to total 2017 population of Finland (5,479 million)

Population of the capital, Helsinki



Population of main summer destinations



Seasonal population conclusions

- Seasonal population requires more data, that is the third operate to participate: market share varies on municipality level.
- Municipality level is enough for Statistics Finland
- It is easy to see how populations differ greatly between weekdays and weekends and especially between the summer holiday peak season and the winter season (out of winter holidays).

Web scraping – Internet as a data source

Statistics Finland

Web scraping – Internet as a data source

- Very much examples especially among European Statistical System: many potential applications
- The most usual target is price statistics (data collection from websites)
- Web-scraping & scanner data for consumer price statistics (2015) was the lead motivator to continue in other statistics at StatFi

Web scraping

- Scrapers are relatively easy to build
- StatFi scrapers haven been built by using open Python packages.
- Service providers scraping data: open social media and open business data
- **Ethics and Big Data: Netiqette**
  - Accept robots.txt, that is a protocol to prevent robots regardless of the national framework and laws.

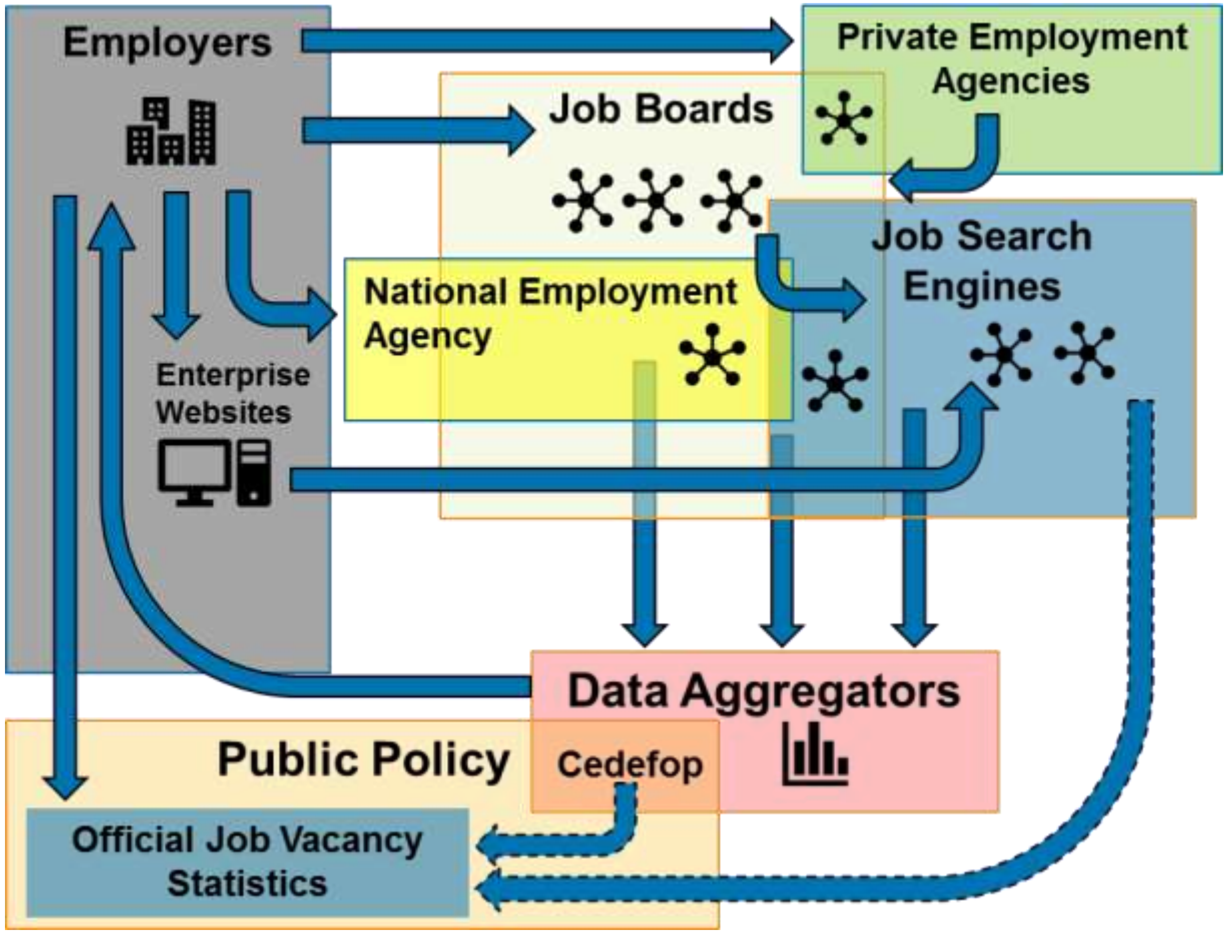
Web scraping: Job Vacancy Statistics

- There are service providers around collecting open data and selling the access.
- First Case Finland: a service provider that scrapes and updates the business data continuously from open platforms.
  - Tests of which one case is Job Vacancy Statistics
- Second Case European Statistical System: project ESSNet Big Data

Web scraping: Job Vacancy Statistics

- Many restrictions and limits
- Obvious target was to collect information from those business that are participating in the official survey.
- Quality of the data
  - Included observations that are not describing an open vacancy but are related to that.
  - Difficulty in defining a single open vacancy among many (scraper collects from many data sources around)
  - Difficulty to get the number of open vacancies
  - Establishment issues
  - In the production there would be too many observations for manual editing.

ESSNet Job Vacancy case conclusions



OJV Data Landscape 2018 by Nigel Swier, ONS, UK.

Job Vacancy web scraping: lesson learned

by Nigel Swier, ONS, UK

- Coverage problems (e.g. not all the vacancies are online)
- No definitive source of OJV data
- Much OJV data is unstructured: text processing and analysis required
- OJV doesn’t necessarily meet the scope of official statistics definitions on a job vacancy.
- A job ad doesn’t correspond directly to the concept of a live job ad (one ad, multiple vacancies)
- ***OJV data is not representative of the labour market and there are definitional issues that make it difficult to compare directly with official statistics***

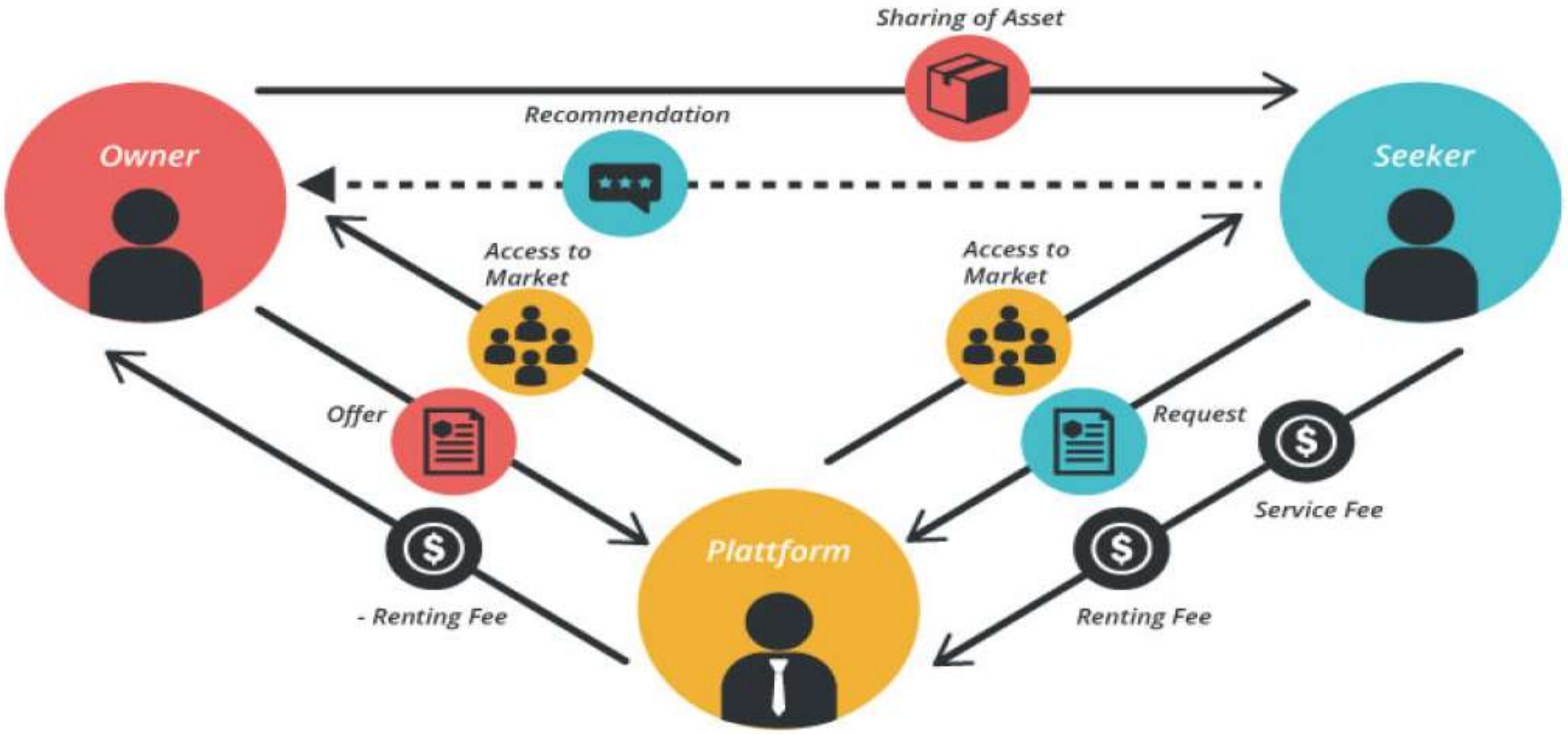
Finnish Job Vacancy case conclusions

- Too messy
- **Make an agreement directly with the open vacancy service providers**.
  - This recommendation holds to many other web scraping potential as well.

Web-scraping holiday homes

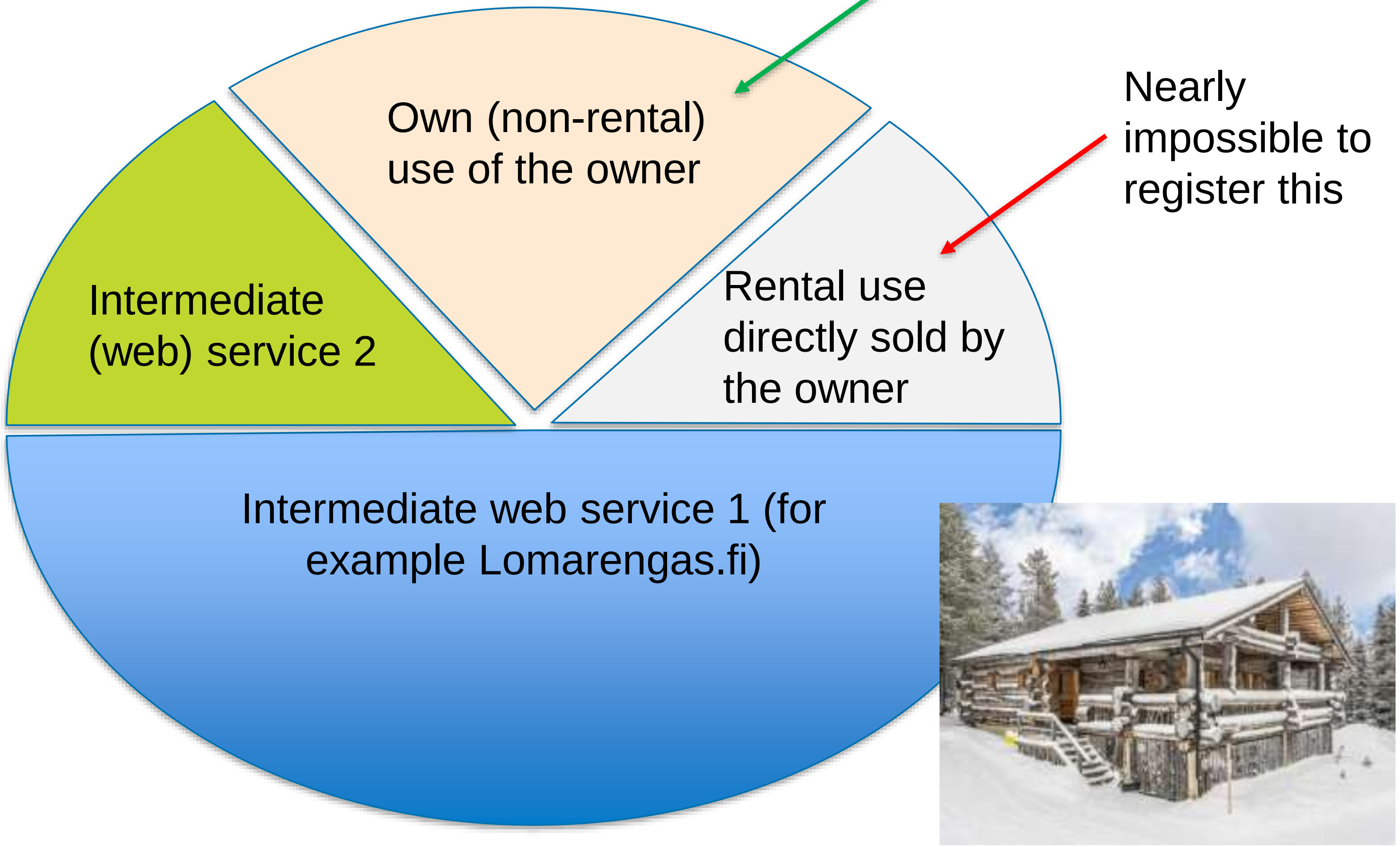
- In Finland, there are roughly half a million buildings classified as holiday homes according to the Finnish Building and Dwelling Register
- Many of these holiday homes / cabins are available for rent on various web platforms
- Accommodation statistics exclude rentals of **privately** owned cabins and holiday homes – a type of sharing economy
- These rentals make up a potentially significant share of total paid accommodation

The sharing economy



Source: Statistics Denmark

The occupancy of a single holiday home throughout the year



Out of scope

Nearly impossible to register this

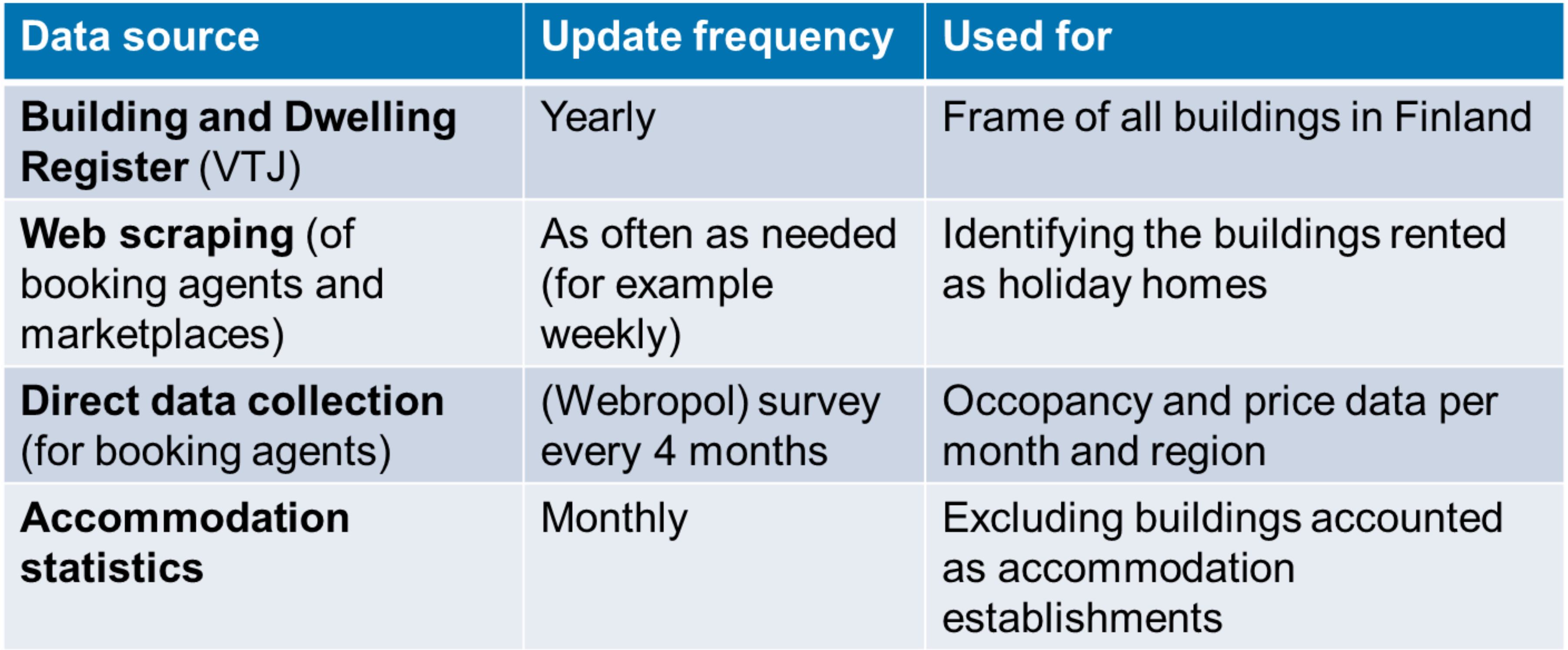
Own (non-rental) use of the owner

Intermediate (web) service 2

Rental use directly sold by the owner

Intermediate web service 1 (for example Lomarengas.fi)

Data sources



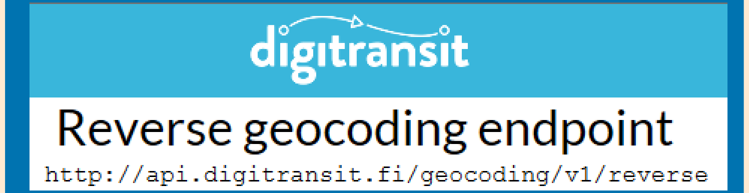
Data source	Update frequency	Used for
Building and Dwelling Register (VTJ)	Yearly	Frame of all buildings in Finland
Web scraping (of booking agents and marketplaces)	As often as needed (for example weekly)	Identifying the buildings rented as holiday homes
Direct data collection (for booking agents)	(Webropol) survey every 4 months	Occopancy and price data per month and region
Accommodation statistics	Monthly	Excluding buildings accounted as accommodation establishments

Web scraping and reverse geocoding

python Web Scraper(s)

XML/JSON

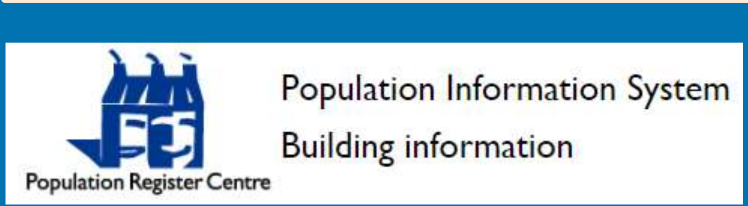
Import to SAS, conversion to sas7bdat format
Data manipulation, for example information retrieval from free text fields etc.



Reverse geocoding endpoint
http://api.digitransit.fi/geocoding/v1/reverse

Address verification
Correct postal codes

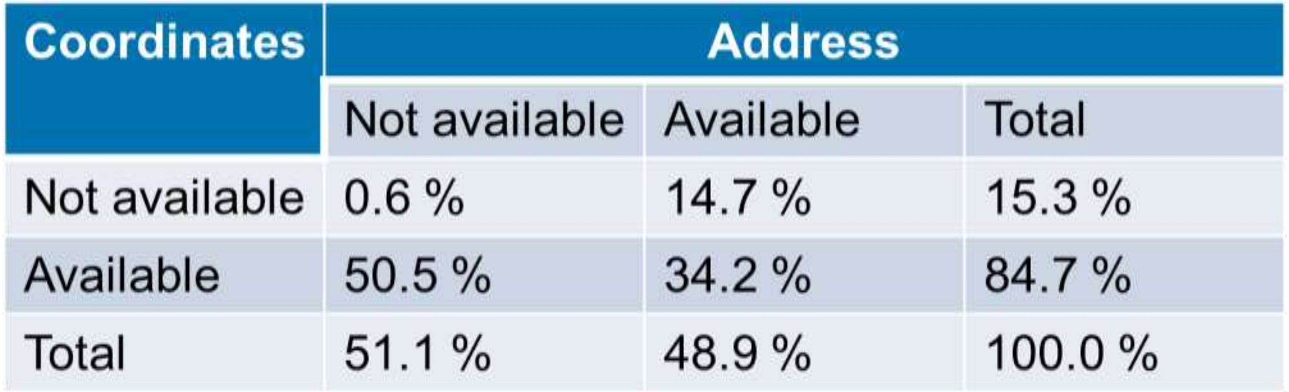
Coordinate reference system chang
From WGS84 to ETRS-TM35FIN



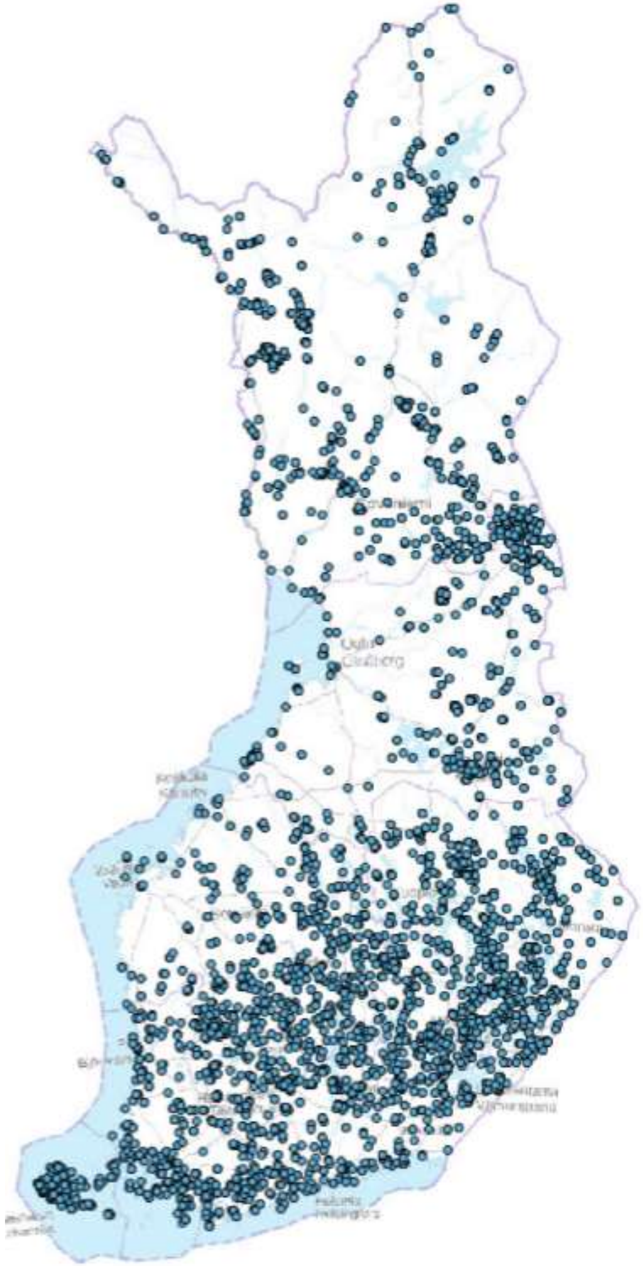
Population Information System
Building information

Nearest coordination points
Closest size match
Closest building year match
Other?

Reverse geocoding results



Coordinates	Address		
	Not available	Available	Total
Not available	0.6 %	14.7 %	15.3 %
Available	50.5 %	34.2 %	84.7 %
Total	51.1 %	48.9 %	100.0 %



Management

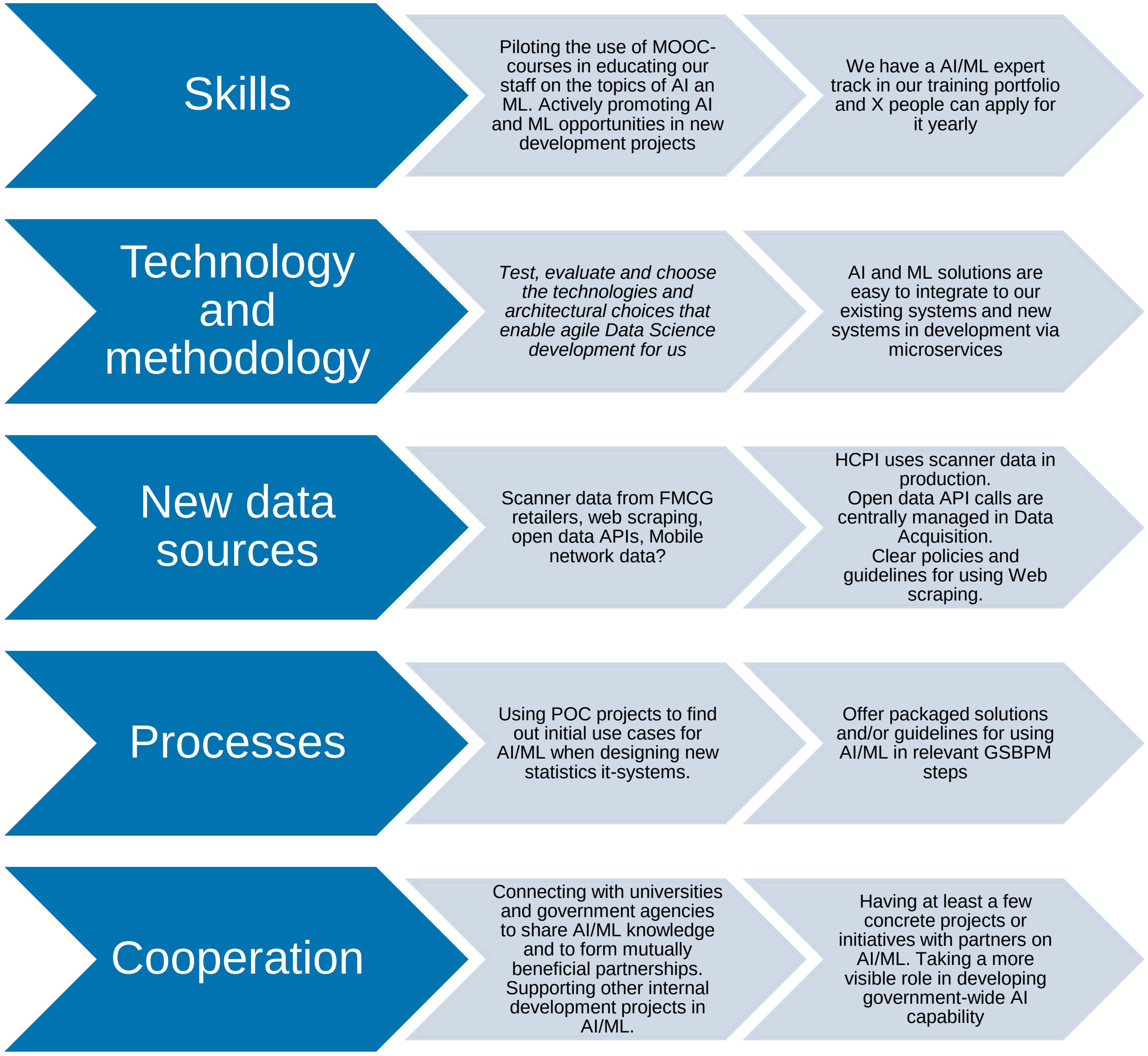
Statistics Finland

New data sources and methods initiative

We aim to

- define the technologies and architectural choices that will enable us to take advantage of the full potential of machine learning and artificial intelligence solutions in official statistics production.
- make it easier for independent dev teams to integrate ML and AI solutions to their products
- educate and encourage dev teams to explore ML and AI opportunities, and to actively consider alternative new data sources (big data, open APIs, …)

Initiative goals



	2018	2019
Skills	Piloting the use of MOOC-courses in educating our staff on the topics of AI an ML. Actively promoting AI and ML opportunities in new development projects	We have a AI/ML expert track in our training portfolio and X people can apply for it yearly
Technology and methodology	*Test, evaluate and choose the technologies and architectural choices that enable agile Data Science development for us*	AI and ML solutions are easy to integrate to our existing systems and new systems in development via microservices
New data sources	Scanner data from FMCG retailers, web scraping, open data APIs, Mobile network data?	HCPI uses scanner data in production. Open data API calls are centrally managed in Data Acquisition. Clear policies and guidelines for using Web scraping.
Processes	Using POC projects to find out initial use cases for AI/ML when designing new statistics it-systems.	Offer packaged solutions and/or guidelines for using AI/ML in relevant GSBPM steps
Cooperation	Connecting with universities and government agencies to share AI/ML knowledge and to form mutually beneficial partnerships. Supporting other internal development projects in AI/ML.	Having at least a few concrete projects or initiatives with partners on AI/ML. Taking a more visible role in developing government-wide AI capability

¡Muchas gracias!

Statistics Finland

Pasi Piela, pasi.piela@stat.fi

Non-traditional data sources in the National Statistical Systems,
17th Meeting of ECLAC, Santiago de Chile