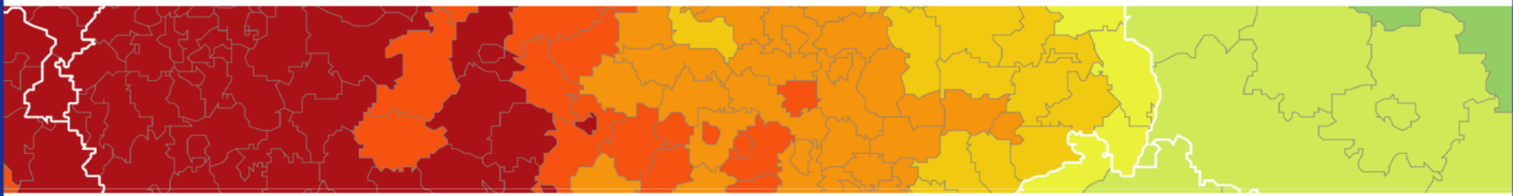
EUROPEAN UNION

Co-financed by the European Regional Development Fund

Inspire policy making by territorial evidence



POTENTIALS OF BIG DATA FOR INTEGRATED TERRITORIAL POLICY DEVELOPMENT IN THE EUROPEAN GROWTH CORRIDORS (BIGDATA)

Targeted Analysis

Final Main Report

28 June 2019

This Targeted analysis activity is conducted within the framework of the ESPON 2020 Cooperation Programme, partly financed by the European Regional Development Fund.

The ESPON EGTC is the Single Beneficiary of the ESPON 2020 Cooperation Programme. The Single Operation within the programme is implemented by the ESPON EGTC and co-financed by the European Regional Development Fund, the EU Member States and the Partner States, Iceland, Liechtenstein, Norway and Switzerland.

This delivery does not necessarily reflect the opinion of the members of the ESPON 2020 Monitoring Committee.

All companies, products, and services named in this report are presented for illustrative purposes only and their mention does not constitute endorsement or recommendation.

AUTHORS

Helka Kalliomäki & Jukka Vahlo
Centre for Collaborative Research, University of Turku (Finland)

Ira Ahokas & Nicolas Balcom Raleigh
Finland Futures Research Centre, University of Turku (Finland)

Jukka Heikkonen, Pekko Lindblom, Ville Keränen & Paavo Nevalainen
The Department of Information Technology, University of Turku (Finland)

Siiri Silm & Anto Aasa
Mobility Lab, University of Tartu (Estonia)

ADVISORY GROUP

Nicolas Rossignol, ESPON EGTC (Luxembourg)
Antti Vasanen, Regional Council of Southwest Finland (Finland)
Dino Keljalic, Region Örebro (Sweden)
Liis Vahter, Ministry of Economic Affairs and Communications (Estonia)

Information on ESPON and its projects can be found on www.espon.eu.

The web site provides the possibility to download and examine the most recent documents produced by finalised and ongoing ESPON projects.

This delivery exists only in an electronic version.

Contact: info@espon.eu

Disclaimer:
This document is a Draft Final Main Report. The information contained herein is subject to change and does not commit the ESPON EGTC and the countries participating in the ESPON 2020 Cooperation Programme. The final version of the report will be published as soon as approved.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF MAPS

LIST OF TABLES

LIST OF BOXES

ABBREVIATIONS

API	Application Programming Interface
AVA	Atomic Visual Actions
CEC	Commission of the European Communities
DL	Deep Learning
EC	European Commission
ESIF	European Structural and Investment Funds
ERDF	European Regional Development Fund
ESPON	European Territorial Observatory Network
EU	European Union
EV	Electric Vehicle
FDI	Foreign Direct Investment
FTA	Finnish Traffic Agency
GSMA	Global System Mobile Association
HICP	Harmonized Index of Consumer Prices
ISO	International Organization for Standardisation
LAM	Liikenteen Automaattinen Mittausasema(Finnish for Traffic Management System, see TMS)
MaaS	Mobility as a Service
NUTS	Nomenclature of Territorial Units for Statistics
NGZ	Northern Growth Zone
O-D matrix	Origin-Source matrix
TEN-T	Trans-European Transport Network
TMS	Traffic Management System (see LAM)

1 INTRODUCTION

This final report of the ESPON targeted analysis, Potentials of Big Data for Integrated Territorial Policy Development in the European Growth Corridors (Big Data & EGC), presents the findings of the one-year-long project that explored how big data can be better utilised to support territorial policy-making in European growth corridors.

Recently the potentials of big data have been emphasised in policy-making due to the fast developments in both the amount of data and the ways it is being handled in policy-making (e.g. Cavanillas et al., 2016; Athey, 2017; Giest, 2017; Cuquet & Fensel, 2018). The combination of the trend of digitizing administrative data, collecting data through diverse devices and rapid developments in data storage has led to the establishment of numerous big data and open data initiatives at diverse governmental scales of policy-making (Giest, 2017), including corridor development (e.g. Harris et al., 2015; Kamel et al., 2016).

Big data is essentially about generating insights from large datasets that in the context of corridor development are related to interesting possibilities regarding spatially, sectorally and temporally integrative policy-making. Irrespective of the many definitions of the term, big data describes broadly the volume and the complexity of the available data, as well as datasets that are too large for traditional processing systems and thus require new technologies (Provost & Fawcett, 2013). For European growth corridors, utilising new data sources such as the SDG Open Data Indicators, could offer numerous benefits for environmentally, economically and socially sustainable policy-making. However, to access this potential, wider data literacy and new kinds of data capacities across all sectors including the public sector are needed (e.g. D'Ignazio & Bhargava, 2015; Koltay, 2015).

After the introduction, where the rationale behind the project as well as the objectives and practical needs of stakeholders are presented, the report proceeds by presenting the conceptual framework of the analysis. After that, the third chapter presents the potentials of big data in general and specifically in integrated policy-making for corridors. The fourth chapter presents the project's three case studies and their findings. In addition, the process and the results of the hackathon that was organised in the project are shortly described. The fifth chapter discusses the implications of the results for corridor governance, and the policy recommendations arising from the analysis. Finally, the key insights from the targeted analysis are summarised in the conclusions.

1.1 GAPS IN THE EVIDENCE-BASE – POTENTIALS OF BIG DATA

European spatial development has for long been about normative envisioning for a more economically competitive, socially cohesive and environmentally sustainable spatial structure (see, e.g., Dühr et al., 2010; Faludi, 2010). However, the integration of diverse sector-based policy approaches has often been completely lacking or superficial (e.g. Stead & Meijers, 2009; van Assche & Djanibekov, 2012; Kalliomäki, 2012; Mäntysalo & Kanninen, 2018). In addition, during the recent years the focus has been shifting towards emphasizing agglomeration economies and urban policies that tend to favour large and dense cities at the expense of peripheral areas (e.g. Puga, 2010; Rodríguez-Pose, 2018). In fact, the increasing concentration of spatial structures as well as increasing societal inequalities (e.g. Kuhn, 2015) have recently directed attention to the value-base of territorial development, which has not been transparent when it comes to the aim of supporting balanced and integrated development (see, e.g., Medeiros, 2019). A stronger evidence-base combined with concrete policy tools are needed to better support both spatially and sectorally integrated policy-making (Chen & Lee, 2018; Rodríguez-Pose, 2018).

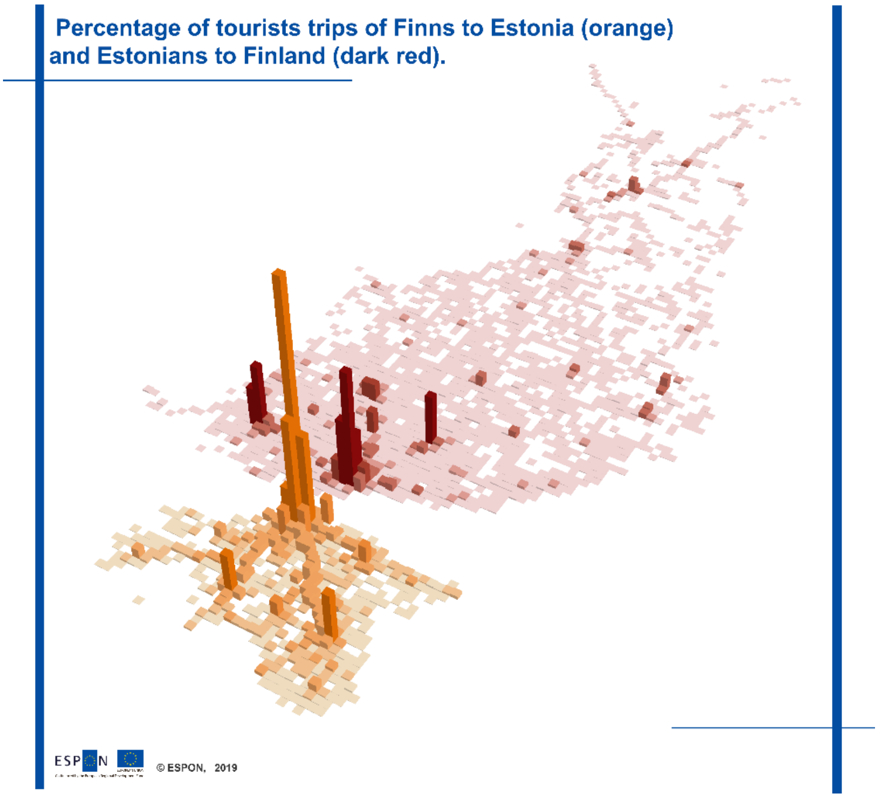
Examples of spatial visions that have combined the urban and rural areas along major infrastructure routes can be found in European growth corridors that cross several administrative boundaries and follow the trans-European transport network (see, e.g., Kalliomäki, 2012; Moilanen, 2012; de Vries & Priemus, 2003; Drewello & Scholl, 2015). Corridor development focuses on increasingly efficient uses of existing infrastructure, which is seen to enable the economic integration of the areas it covers and sustainable transportation. It aims to promote spatial and transportation planning based on a connecting infrastructure. Growth corridors also function as examples of opportunities for spatially and sectorally integrative policy-making that combines diverse sectoral approaches and supports territorial integration.

However, the abstract nature of corridor-based spatial visions have been criticized for their inability to engage with regional development practices at the local and regional level (e.g. Kalliomäki, 2012; see also Davoudi 2009). Planning for functional, cross-border corridors can be challenging due to the need to coordinate among various territorial administrations (Moilanen, 2012; see also Faludi, 2018). Furthermore, the utilisation of spatial data that adequately captures the functionalities of corridors related, for example, to daily mobilities and practices has been so far rather weak in planning processes, especially in cross-border areas (Kotilainen et al., 2016; Silm & Tiru, 2018).

Very often the absence of data is the main limiting factor that prevents high-quality estimates of cross-border mobility patterns. In this project, insights regarding the regular mobility of people in Estonia are produced with the help of mobile positioning data (see Case 3 in Chapter 4.3). This same mobile positioning data can also be used to produce insights into cross-border movements as aggregated mobility flows. Such cross-border movements can be assessed based on roaming data (incoming and outgoing). Data from mobile operator(s) in Estonia can describe inbound flows of people from other countries visiting Estonia, and outbound flows of people from Estonia to other nations. If it were possible to access, combine, and analyse data from mobile network operators in several nations, then the volume of mobility flows through European corridors could be estimated on this basis and those insights could be used by many administrative bodies to make evidence-based policy.

Mobile positioning data also supports the measurement mobility flows in cross-border functional regions and twin cities (Figure 1.1). In addition to the volume of movement flows, mobile positioning data allows segmentation of the flows based on trip characteristics such as frequency and duration. For example, tourists, transnationals, commuters, long-term stayers can be distinguished using such variables (Ahas, Silm & Tiru, 2017).

Figure 1.1. Percentage of tourists trips of Finns to Estonia (orange) and Estonians to Finland (dark red).



When European growth corridors are thought of as integrative frameworks for meta-governing spatial development, evidence derived from analysing big data can improve the legitimacy of corridor-based policies and planning processes. The role of data is central in understanding complex, often place-based development challenges. Understanding both micro- and macro-level dynamics and mobilities becomes increasingly important in transnational corridor development (e.g. Gutiérrez et al., 2011). In the end, corridor development policies intending to promote sustainability aim at affecting behavioural patterns of various societal actors such as individual citizens and companies (see, e.g. Enoch, 2012; Zijlstraa & Vanoutrive, 2018). For this reason, the need to better understand the motivations and drivers of physical mobility highlights the significance of developing resources useful to more comprehensively examine of spatial connectivities. At its best, utilising big and open data can help in creating new businesses, supporting cross-border services, enhancing resource efficiency and sustainability, as well as increasing participation and transparency of decision making. For example, smart mobility initiatives (e.g. Mobility as a Service) and automated mobility pilots (e.g. the 5G cross-border pilots, see EC Digital Single Market, 2018) can boost the functionality of growth corridors, create new business opportunites, and support sustainability.

Nowadays, the arguments about functional corridors are often based on strategic objectives that lean on an assumed importance of infrastructure expansion without an adequate evidence-base (e.g. Kalliomäki, 2012; Kotilainen et al., 2016). Furthermore, infrastructure projects often have been used as tools of competition and cohesion policies without adequate understanding about the suitability of these approaches to actual place-based needs (see, e.g., Rodríguez-Pose, 2018). The knowledge-base used in corridor development has leaned mostly on statistical data related to administrative areas at diverse scales. The functionalities and flows along corridors also have been described mostly between the administrative areas. (Kotilainen et al., 2016.) Overall, the dynamics of urbanisation are poorly understood due to the lack, and insufficient use, of new data sources (Caprotti et al., 2017). This is a major knowledge gap for growth corridors that should be addressed in the analyses of changing territorial dynamics.

Altogether, there are currently few detailed analyses about the functionality of growth corridors, highlighting the potential of big data in increasing the understanding about connectivities along the corridors. Concerning the analyses in the projects of ESPON program, most functionality-related research has focused on urban areas, and the analyses of flows and interactions have focused on accessibility and physical flows (e.g. Nordregio, 2005; Spiekermann & Wegener, 2007; Spiekermann et al., 2015). More recently, some analysis has focused on the territorial impacts of refugee flows and global Foreign Direct Investment (FDI) flows toward Europe, as well as on the territorial dimension of cross-border public services and the transition towards a circular economy.[1]

1.2 OBJECTIVES OF TARGETED ANALYSIS

The aim of the Potentials of Big Data for Integrated Territorial Policy Development in the European Growth Corridors (BIGDATA) targeted analysis project is to research the potentials of big data to better inform comprehensive territorial policy in European growth corridors. In particular, the objective is to find and evaluate new available data sources for evidence-based policy-making regarding the Northern Growth Zone (NGZ) in the Baltic Sea Region.[2] The primary interest is thus in analysing new datasets that can describe diverse flows and interactions along the corridors and that can produce new insights regarding functional corridors and integrated territorial policy development. Therefore, the project aims at broadening the knowledge-base of corridor development that has so far mostly leaned on datasets describing physical flows of people and goods.

By examining big data-related potentials for policy-making, the aim of this targeted analysis is to address the needs of project stakeholders in Finland (Regional Council of Southwest Finland), Sweden (Region Örebro) and Estonia (Ministry of Economic Affairs and Communications). The project's four cases seek to generate new knowledge about the functionalities of the NGZ through processes that could be replicated in other European growth corridors. In this way, the project contributes to Specific Objective 2 of the ESPON 2020 Cooperation Programme: supporting stakeholders in utilising territorial evidence in their policy development.

The geographical context of the targeted analysis is the NGZ, which mostly follows the TEN-T corridors in the Baltic Sea Region. In the corridor's international strategies, the corridor stretches

[1] See https://www.espon.eu/Flows-Interaction (Accessed 10 May 2019).
[2] See http://www.turku.fi/en/northern-growth-zone (Accessed 8 February 2019).

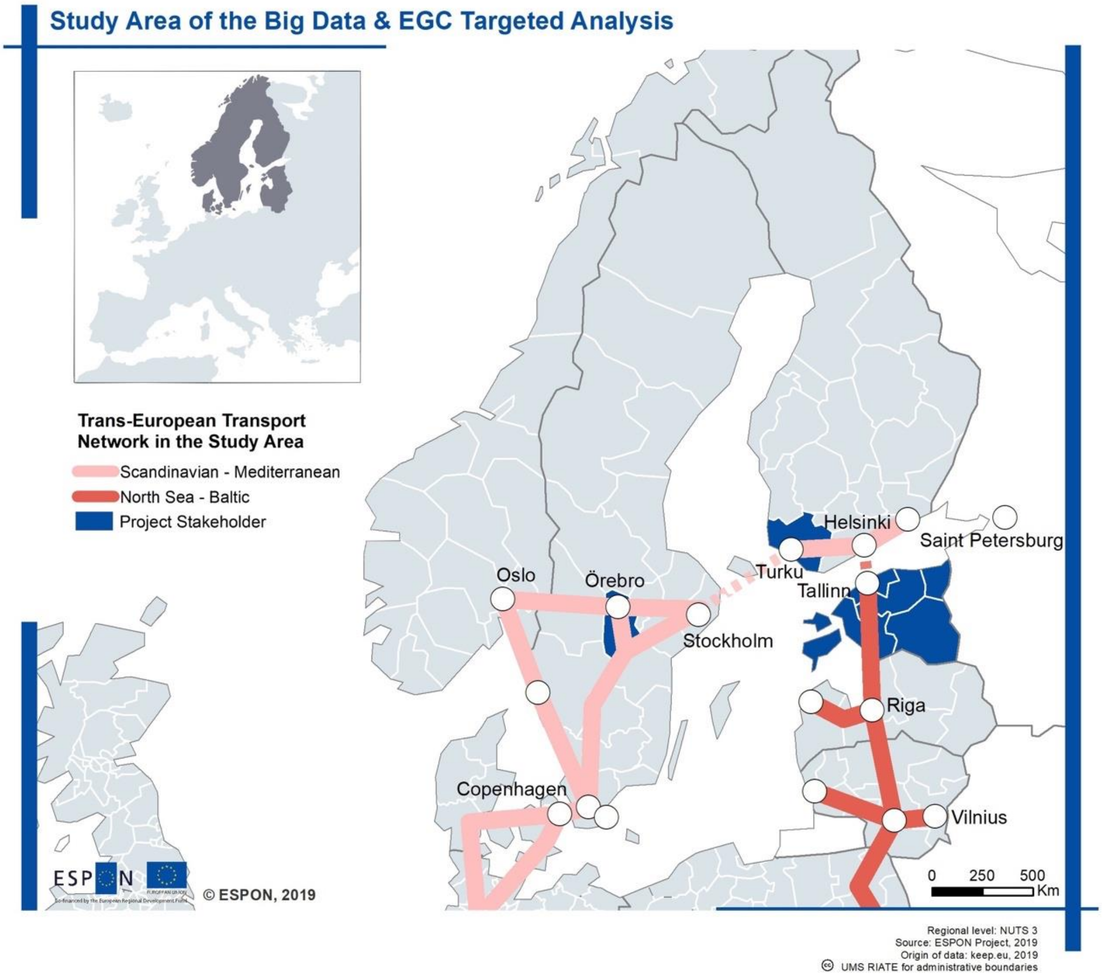
from Oslo – via Örebro, Stockholm, Turku and Helsinki – to St. Petersburg. In addition, corridor-based collaboration exists between Finland and Sweden, most recent strategy being the Talsinki twin-city initiative. More specifically, the targeted analysis is focused on the area covered by the three stakeholder territories: Finland, Sweden and Estonia (Map 1.1).

The NGZ aims at boosting the region's competitiveness in the global arena through the creation of a single, internationally recognized market and commuter belt. However, a central challenge in the NGZ is that it has been developed through partial solutions and organizational silos that are not optimally serving the development of the corridor as a whole. For example, transportation planning has long been too separate from land use planning even though the location of jobs and housing is a crucial factor for sustainable mobility (cf. Kalliomäki, 2012). In its planning and development processes, the corridor is still mostly leaning on more traditional data sources (Kotilainen et al., 2016; Santala, 2017), although recently there have been diverse projects exploring the possibilities for new approaches (e.g. Kalliomäki et al., 2018; Silm & Tiru, 2018). For example a project exploring the connectedness of maritime company board members in the NGZ utilised Finnish Asiakastieto portal[3] as a data source (Kalliomäki et al. 2018).

The data and methods used in this targeted analysis are versatile. While the methodology related to each case study is described more in Chapter 4, the overall approach is summarized here. A literature review was conducted to ascertain the state of the art in big data and integrated policy-making. The data portals run by key organisations in the study area were searched and reviewed to gain an overall understanding about available datasets and their status. Company-led (e.g. MaaS Global) and industry-led (e.g. ISO) efforts to standardize data practices were also identified. In addition, four interviews were made with staff from various public organisations representing municipal, regional, and national perspectives to understand the various strategic and operational aspects related to utilising big data in the public sector, particularly in the context of territorial development (see Appendix 1). These interviews were conducted to gain a practical perspective on the issues arising from the literature review and they are used as testimonials throughout this report. Furthermore, environmental scanning regarding data governance and big data driven futures was conducted based on recent policy reports and extant literature. This environmental scanning serves the basis for this report's policy recommendations for both strategic and operative levels of corridor development.

[3] https://www.asiakastieto.fi/web/en/

Map 1.1: Study Area of Big Data and European Growth Corridor Targeted Analysis



1.3 PRACTICAL NEEDS OF STAKEHOLDERS

The targeted analysis is based on the practical needs of the three project stakeholders: The Regional Council of Southwest Finland (lead stakeholder), the Region of Örebro in Sweden, and the Ministry of Economic Affairs and Communications in Estonia. During a project workshop, the stakeholders identified three themes as the most important policy dimensions related to corridor development that would benefit from big data: 1) infrastructure and connectivity planning; 2) regional economic development, and; 3) land-use planning (Box 1.1). These themes have been also recognized in the literature as three main areas of corridor development as corridors function as axes for 1) transportation and logistics, 2) economic development, and 3) urbanisation (Zonneveld & Trip, 2003).

Box 1.1: Three Policy Themes Identified by Project Stakeholders

Infrastructure and connectivity planning. Considering infrastructure, stakeholders stressed that promotion of connectivity and investments in transportation infrastructure is an important policy area in territorial development that could significantly benefit from big data. The investments can be more efficiently directed based on more accurate knowledge -- e.g. about the flows of people and goods -- which are at the core of corridor development. Especially, mobility data is needed at the corridor level to support integrated policy-making combining mobility with economic development, housing and service provision. Furthermore, better understanding of corridor functionalities supports the planning of fast train connections.

Economic development. Policies related to economic development, and particularly new business and job creation, could benefit greatly from big data. New types of data are needed to support regional economic development, enhancing partnerships with private companies and more efficient allocation of investments. Knowledge about different territorial networks related to, for example, organizational interlinkages and project networks, as well as interregional collaboration in education and between companies as well as research and development would be useful for actors at various organisations and scales. In addition, development of education and culture contribute to economic development. For instance, knowledge on the mobility of students in the corridor would be useful for planning collaboration in higher education.

Land use planning. Land use is another policy area that could benefit substantially from big data. Knowledge about the mobility of people, information and goods supports this policy context as well. In addition, safety related policies, as well as environmental sustainability policies were brought out as policies that could be renewed based on new insights. Those policies are strongly interlinked with land use and infrastructure promoting connectivity.

The themes identified by the stakeholders reflect the strategic objectives of major growth corridors in the study area: transportation, smart mobility and commuting as well as developing the corridors as platforms for innovation. Even though big data utilisation was not mentioned directly as a key strategic objective in the study area, this project's interviews support the view that utilising big data is built into most corridor development practices as an assumed positive development. These interviews suggest that increasingly efficient utilisation of data can play a central role in advancing smooth mobility and flows along the corridors, as well as in the development of uniform labour market areas and (digital) innovation platforms. Furthermore, the need for new sources of data describing the interactions and connectivity along the corridors was mentioned as evident in the practical attempts to improve the data-driven decision-making in corridor development.

Even though the themes identified by the stakeholders represent different policy sectors, they are also strongly interlinked in territorial development practices. This interconnectedness is especially true in corridor development that utilises connecting infrastructure as a framework for promoting corridor functionalities. There is a need for an integrated and collaborative approach to data-driven corridor governance in order to create comprehensive understanding about corridors' development dynamics which can inform policy-making (cf. Chen & Lee, 2018).

The three themes identified in the needs analysis were taken as the basis for data categorisation concerning corridor development while taking into account their diffences and various interrelations. As one aim of the project was to identify datasets describing the interactions and flows along the growth corridors, the category of land use presents a rather static framework for flows and interactions happening in relation to the other identified policy categories. However, land-use planning in the context of growth corridors essentially benefits from big data describing e.g. transportation flows and economic interactions hence highlighting the need for an integrated approach.

In this project's inception report, an extensive overview was presented about the stakeholders' public policies and data related needs. Central themes were related to better governance and data driven innovations, and differing needs between the public and private sector. Especially the different processes and varying timescales for these sectors were discussed. In the private sector big data is typically used for short-term decision-making, whereas in the public sector, democratic decisions and planning processes usually imply longer time horizons (e.g. Kim et al., 2014). However, interesting potentials concerning new forms of big data and public policy-making are partly related to the possibilities in utilising recent data to speed the production of provisional indicators for policy makers. Such potentials require increased remote monitoring capacities and access to real-time data to make short-term forecasts and revise long-term ones. The motivation for such efforts is to give decision-makers more up-to-date information about the present situation to enable more timely planning of development activities. However, tapping into these potentials requires appropriate data governance policies, shared standards for using data, advanced analytics expertise, and wider data literacy in public sector organizations and broader society.

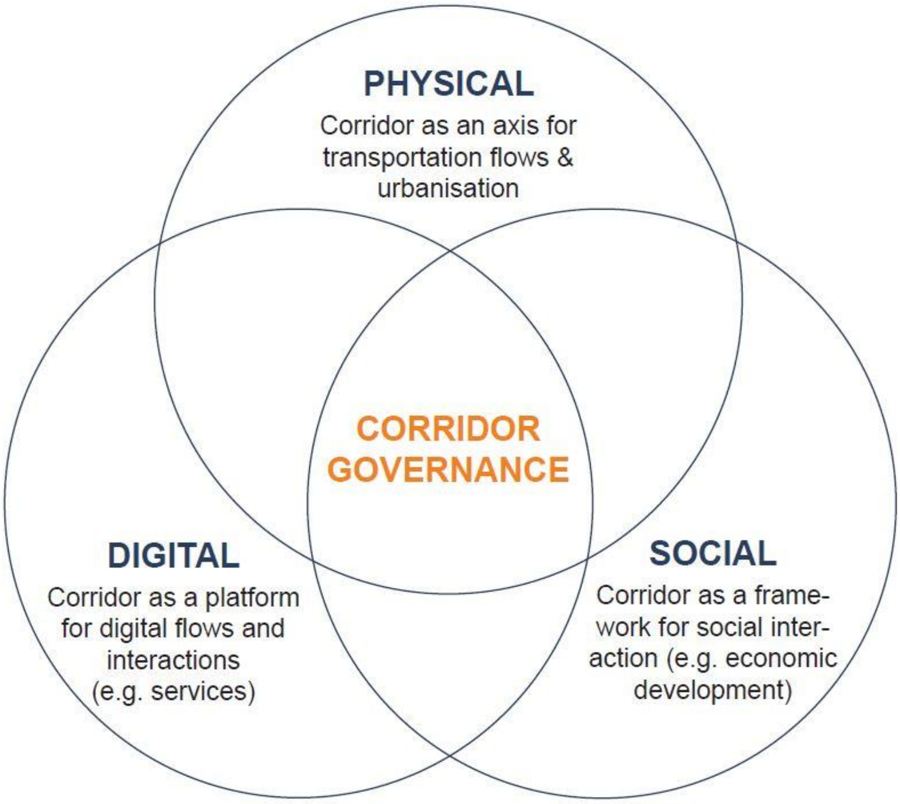
2 CONCEPTUAL FRAMEWORK

The approach for identifying new data sources is built on a conceptualisation of corridors as meta-governance frameworks for territorial development (see, e.g., Zonneveld & Trip, 2003; Jauhiainen & Moilanen, 2011; Moilanen, 2012; Kalliomäki, 2012). Corridors as meta-governance frameworks aim at integrating diverse development practices sectorally, spatially, and between multiple actors involved in spatial development thus forming a basis for collaborative governance and an integrated approach to data utilisation. For example, corridors between cities aim to facilitate the smooth flows of labour and goods by improving transportation infrastructures and services and by means of sustainable land use planning. Today's challenges related to climate change and accelerating urbanisation have increased the importance of urban growth management and spatial planning at higher planning scales, including the national/state level, where inter-sectoral and spatial coordination of urban growth management issues have become more generalised (e.g. Kalliomäki, 2015). For many nations, these factors have increased the importance of corridors as a new frame for territorial development.

A conceptual framework was developed in the project to support the identification of datasets with potential for comprehensively describing flows and interactions along European growth corridors. The framework builds on the geographical conceptualization of space as absolute, relative and relational (Harvey, 1973), which here describe physical, social, and digital aspects of corridor development (e.g. transportation flows, social and intellectual networks, or services). The approach presents a way to operationalize a theoretical conceptualization of space as three aspects of corridor development which are often interrelated and overlapping (Figure 2.1; see Jauhiainen & Moilanen, 2011). For example, land-use planning has for long been conducted in the domain of absolute space, in which land forms "a surface" for different uses and bounded territories (cf. Faludi, 2018). Relative and relational views change the perspective of planning as both social and digital realities affect the ways in which the physical space is organised, which comes out for example in the new working arrangements made possible by digitalisation.

The purpose of the framework is to broaden the perspective of corridor development by supporting inclusion of physical, social and digital dimensions in policy-making considerations. In this targeted analysis, the proposed framework is tested through three case studies exploring fluctuation levels of national transport and commuting; project network analysis; and passive mobile positioning data for transport infrastructure analysis. These cases introduce new pathways for analysing, integrating and utilising big data to widen the horizon about new functional geographies. They show that in addition to sectorally and spatially integrative policy-making, integration of various datasets can support a more comprehensive understanding about changing territorial dynamics.

Figure 2.1. Conceptual framework describing the three overlapping spatial dimensions of corridor development that should be taken into account in integrated territorial policy and corridor governance



The framework is intended to illuminate multifaceted characteristics of flows and interactions relevant to corridor development. It conveys the the importance of seeking variety in the sourcing, production and management of data for analysis in the pursuit of comprehensive corridor governance. In seeking to understand the complex interactive nature of functional growth corridors, there is a need to combine not only different datasets describing their diverse characteristics but also conventional and unconventional (or new) datasets. These data-driven insights can better ground the democratic decision-making structures and planning processes of corridor development based on more comprehensive insights. Because today many interrelated geographic functions are largely driven by, publicly understood through, and mediated by digital technology, understanding multiple spatial dimensions of corridors related to physical, social and digital spaces supports a more comprehensive framing of policies and the efficient utilisation of research results (cf. Ash et al., 2018). For example, the results of the mobile positioning research in Case 1 have different implications on developing the corridor as a physical space (e.g. infrastructure planning), as a social space (e.g. balanced development, economic development) and as a digital space (e.g. service development), that, however, should be reviewed as a whole for efficient coordination of sector policies and thus allocation of resources and prioritizing development activities.

The functionality of a growth corridor basically consists of spatial interactions, which “naturally form a network/graph, where each node is a location (or area) and each link is an interaction between two nodes (locations)" (Guo, 2009, 1041). Whereas the term flow refers commonly to unidirectional flow from one location to another (e.g. a number of migrants, flows of goods), the term interaction normally refers to two-way, reciprocal action between people or things (see, e.g., Spiekermann & Wegener, 2007). From the perspective of functional corridors being characterised by numerous interactions and flows, they can be seen as a complex assemblage consisting of multiple functionalities and connectivities (see, e.g., De Landa, 2006).

The definition of a functional corridor has connections for example to the debates on city-regions, Functional Urban Regions (FURs) and polycentric urban regions (e.g. Kloosterman & Musterd, 2001; Dühr, 2005; Meijers, 2005), to name some of the key concepts dominating discussions in the 2000s. In fact, while understanding of FURs has normally been built on the conception of the city-region as a functional economic space and as a top-down framework of urban governance delineating traditional administrative boundaries (Davoudi, 2008), the attempt to understand the multiple functionalities at the level of inter-urban and transnational growth corridors extends the analysis to wider spatial context that builds on the scalar and sectoral complexity of these fluid spatial assemblages (see Jauhiainen & Moilanen, 2011). Corridors, when seen as meta-governance frames of spatial development, bring together multiple scales and sectors. Their development has to be approached from a relational perspective that, however, also necessitates certain spatio-temporal fixes that represent themselves through multilevel governance. (cf. Jessop, 2006; Jones, 2009.) A corridor itself can be seen as a strategic layer of governance building on a spatially and temporally grounded understanding of a functional corridor. However, as the multiple functionalities of a corridor are in a constant flux, the fluid nature of corridors should be recognized in the practices of strategy setting and policy-making.

Traditionally, territorial policy-making in growth corridors has mostly utilised data describing the physical flows in space (i.e. transportation and commuting data), whereas the digital and social aspects of corridor development have been relying more on assumptions without a detailed enough knowledge-base. However, new datasets that describe digital and social interactions hold potential to generate new forms of insights into spatial connectivities and interdependencies among corridor actors. For example, an analysis of maritime cluster connectivities in the NGZ brought out the strong interlinkages between the NGZ and Åland, even though Åland is not officially part of the NGZ network (Kalliomäki et al., 2018). An additional resource that is largely untapped is the anonymized caller-called partners in mobile positioning data which can describe corridor functionalities in terms of social networks rather than strict physical movements. These kinds of new insights from big data seem particularly promising for developing this corridor as a platform for multiple connectivities and explaining some of the *why* behind physical movements of people, vehicles, and material as well as non-physical flows of information and knowledge in space and through time.

3 BIG DATA AND INTEGRATED TERRITORIAL DEVELOPMENT

3.1 SITUATING INTEGRATED TERRITORIAL DEVELOPMENT IN THE BIG DATA LANDSCAPE

Data is continually created, distributed, stored, and utilized across all sectors, at all levels, and at all scales in daily life. This data is often granular, featuring many observations of specific entities at fine spatial and temporal resolutions. In part, this granularity is driven by individual actors continually producing data — or having data produced about them — via their many digitally mediated interactions with other people, sensors, computing devices, organizations, and administrative bodies. In this data-rich setting, it is now possible for policymakers to use this data to make evidence-based integrated policy for territorial development.

Data users, organizations, information technology systems, and devices can be conceptualized as a dynamically changing ecosystem in which interconnecting needs of organizational and individual actors drive the development and consumption of technology to produce, store and use data to achieve specific goals. This evolving and intertwined set of goals act as push- and pull-drivers which influence what data is sought, produced, kept, or taken into use over time. For any given future time, the forms of data circulating in a data ecosystem will be contoured by the past needs and goals of actors and their assumptions regarding what data could be made to exist and its perceived value for the future. In assessing the value of any given data source, policymakers and data scientists need to take into account broader contexts and specific questions they seek to answer now as well as the kinds of questions that may need answering in the future. Making these valuations will require improved data literacy combined with an active appreciation for potential developments of new and newly available data and how it can be analysed.

The concept of big data originates from the needs of first major Internet companies and refers to the systems needed to store, manage, access, and explore large datasets. Since the term was coined by Weiss & Indurkhya (1998), Diebold (2000), and Laney (2001), it has steadily grown in popularity and use. Despite its hype, the main premise of big data continues to be relevant: new insights about how things happen can be produced by combining and analysing large, finely detailed datasets. The pursuit of these new insights motivates the defining activity in the big data landscape for 2018 and 2019—the rapid development and deployment of advanced tools for data analytics coupled with increasing demand for larger datasets. Big data and AI-based analytics will become increasingly common in all forms of organizations in the coming years. Ethical standards for using these new tools are also co-evolving as more public and private organisations begin routinely using big data to inform their actions.

Several trends combine to make big data a significant source of insights for territorial development. First, the trend of digitalization, starting in the 1990s and today reaching nearly every sector of society still has much more to offer in terms of producing efficiencies, transparency, and decision-support to governments and industry leaders. Many governments, companies, and industries are actively deepening the digitalization of their services. Interoperability standards continue to develop serving an important role in making common practices across industries and sectors that can produce high-value. Second, new advanced data analytic tools such as machine learning, Artificial Intelligence, neural networks, and deep learning are becoming more widely available due to advances in information technology and growth in algorithm exchanges (e.g. IBM's Model Asset eXchange[4]). And third, the value of sharing and exchanging data is now broadly appreciated by public and private organizations. The availability of data via portals and APIs reduce access and extraction costs for many kinds of datasets. As digitalization continues to expand, more datasets with finer grain resolutions, covering more potential use cases, can be expected.

3.1.1 New Advanced Analytic Tools for Generating Insights

In the near future, an avalanche of widely available and mainstream advanced tools for data analytics will likely open up new ways to produce insights from big data. As of the beginning of 2019, many such tools and services are already on offer to organisations from long-standing companies and start-ups. Many of these new tools and commercial services (for examples, see Appendix 3) aim to simplify the workflow of data analytics for end users, others expand the analytics toolkit for expert data scientists, and still others aim to tackle difficult ethical challenges for companies such as detecting algorithm bias (e.g. IBM's AI Fairness 360 toolkit[5], see Feldman 2018b; or DSAPP's Aequitas[6]) or setting principles for AI development (e.g. Atonium EISMD AI4People Initiative[7]). In addition to tools and services, neural network exchanges are now well-established – reducing how often the tedious work of training new neural networks to perform specific tasks (e.g. object detection in images) needs to be repeated and enabling the rapid deployment of these algorithms to new settings. The landscape of available tools for big data is continually evolving and is largely driven by business and startup activity. FirstMark Capital has been tracking this 'scene' since 2012 and publishes an annual big data Landscape which illustrates tools and their application areas (Figure 3.1).

---

[4]. https://developer.ibm.com/exchanges/models/ (Accessed 1 February 2019).
[5] https://developer.ibm.com/code/open/projects/ai-fairness-360/ (accessed 6 February 2019).
[6] https://dsapp.uchicago.edu/projects/aequitas/ (accessed 6 February 2019).
[7] http://www.eismd.eu/wp-content/uploads/2018/11/Ethical-Framework-for-a-good-AI-Society.pdf (accessed 6 February 2019.)

Figure 3.1: Categories and subcategories of big data tools from FirstMarks Capital's Big Data and Artificial Intelligence Landscape. Source: Turck & Obayomi (2018).

The 2018 landscape depicts four main areas of activity: infrastructure and analytics; applications targeted to enterprises or industries; tools, software, coordination, and cross-fertilisation from open source activities; and data sources for various kinds of data objects and data resources such as services, schools, and research. As new analytic tools and infrastructure become available as customized applications for specific sectors and task categories, new data sources of the thematic categories of health, economics, transport, people, vehicles, locations, home assistants and IoT, sensors, and others will be increasingly available to produce new data-driven services. As this landscape develops, new tools and practices will become common in both private and public organizations fueling greater interest in their use. Additionally, these new tools will likely reduce barriers in many sectors to producing useable insights. Now may very well be a moment for AI-driven data analytics similar to the moment before email became ubiquitous in most workplaces – such systems could soon become a banal part of daily working life. These changes present uncertainty and vast opportunity to policymakers, their organizations, and the people they serve.

3.1.2 Identifying New Data

A key pathway to generating new insights is to use new datasets. However, what can be considered as new or unconventional is a matter of perspective. From a public sector policymaker's point of view, datasets made available by the private sector could be seen as new, while many familiar public sector datasets seem old. The opposite can be true for a private sector actor. Of familiar 'old datasets', a form of newness can be introduced by the accumulated depth of the temporal dimension (e.g. data from Google Maps Streetview, FlickR, Twitter, or a national population register). Newness also comes from novel combinations of familiar datasets aimed at producing richer insights into phenomena. For this purpose, the conceptual model proposed in chapter 2 can serve as a useful guide. Combining data can also help achieve new forms of multi-scale analysis. EU-level, national level, and municipality-level open data portals contain many datasets generated by public sector organizations which can be matched across municipalities, regions, or nations. This is becoming increasingly possible as EU member states continue to develop their open data portals and follow best practices for open data.[8]

Kanellos' (2016) proposes a few concepts to help identify new data sources. These include Fast Data, 'good enough' forecasts rapidly produced from large datasets for instantaneous use (e.g. guessing a stock's price a few hours from now); Dark Data, data that exists but is not easy to access because it is un-networked, unstructured, or in unusual formats (e.g. security video); Lost Data, data generated in industrial and commercial operations but locked away in proprietary systems (e.g. an oil drilling platform has 30,000 sensors but data from only 1% are ever utilized by data scientists); and New Data, 'data we could get and want to get but aren't harvesting now' – data that will require new data products and services. This last concept links new data creation with actor intentions to produce, store, and use new kinds of data as a motivator for creating or implementing the systems required to do so.

New forms of data also emerge over time as new sensor technologies are developed and become popular (e.g. 3D cameras for mobile devices, see Kelly 2018), networks of sensors are installed in cities and rural landscapes (e.g. air quality monitoring for cities or soil moisture monitoring for agriculture), and data-driven private sector products and services are launched (e.g. platforms for business intelligence or cloud-based AI analytics tools). This process of new sensor technology

[8] European Open Portal Maturity Index, https://www.europeandataportal.eu/en/dashboard#2018 (Accessed 31 January 2019).

emerging and becoming widely distributed is already evidenced by the already widespread distribution of GPS, Wi-Fi, Bluetooth, gyroscope, accelerometer, compass, light sensors, sound sensors, microphones, cameras, barometers, thermometers, and hygrometers found in most smart phones. Developments in new sensor technology are a valuable place to look for new and emerging data sources. For example, traffic sensors utilising LiDAR, laser arrays that map 3D representations of motor vehicles, bicycles, and pedestrians, are now sold as a kind of smart city solution[9]; and cube satellites (cubesats) – the size of a loaf of bread – combined with increasingly available commercial space launch services give regional authorities and companies affordable access to dedicated space-based monitoring with fine-grain temporal and spatial resolution.[10] Many other kinds of remote sensing systems are already on the market and more – some of which are difficult to imagine today – can be expected in the future. All of these emerging sensors will become new sources of data and will introduce new forms and resolutions of data.

New data sources become available as they co-evolve with new tools and services created to meet user demands. Companies are offering new kinds of data-driven services which simultaneously open new capacities for organizations and drive demand for data and timely insights. For example, several companies sell space-based monitoring services (Broad, 2018). The company Planet Lab operates a large constellation of more than 340 cubesats and offers businesses, governments, and others visual space-based monitoring at a spatial resolution as fine as 72 cm with daily updates.[11] The startup Orbital Insight analyses satellite images to generate economic insights into the retail activity of specific companies by the proxy of activity in 260,000 retail parking lots (Metz, 2019). Back on earth, there are companies are producing data-driven insights from land-based data sources about customer behaviour. The company UNA provides granular tracking of entity and individual movements, synthesizing many kinds of data sources, to produce a “Real World Graph”. Their business model incentivises data owners to sell their data, industrial sectors to purchase marketing insights, and retailers to commission fine-detailed reports about consumer behaviour.[12] Mobile phone operator Telia (data provider in Case 3, see chapter 4.3) offers its ‘crowd insights’ service which includes journey and footfall analysis for municipalities and regional governments based on the company’s user data.[13]

Data-driven services are also being developed for supply chains and shipping. For example, the company TradeLens, a partnership of IBM and Maersk, aims to digitalize the shipping industry using blockchain and smart contracts to generate supply chain efficiency through real-time data sharing (IBM, 2018). In addition to the TradeLens initiative, there are long-standing efforts to digitalise shipping documents for air freight through eAWB (see Sauv, 2018) and cargo by sea. Such, fully

---

[9] For example, one smart city vendor sells a sensor that integrates LiDAR object detection and a camera for use in Traffic Management Systems. This sensor can, for instance, determine how many people are standing at an intersection waiting to cross a street or the precise length of a vehicle that passes.

[10] For a comprehensive list of all cubesats see https://www.nanosats.eu/ (Accessed 24 January 2019).

[11] https://planet.com (Accessed 24 January 2019).

[12] https://www.unacast.com/products, (Accessed 22.1.2019).

[13] https://www.teliacompany.com/en/about-the-company/internet-of-things/crowd-insights/, accessed 24.1.2019 and Sharma (2018).

digitalised cargo and freight manifests would open up new possibilities for economic indicators by introducing a means for creating big data-driven market signals that are faster than price proposed by Mayer-Schönberger (Oxford Internet Institute, 2018; Mayer-Schönberger & Ramge, 2018). All of these examples show how new data become available as services and products are developed to meet market demands.

While the past few years has placed much attention on the value of public open data, private-sector companies have found good reasons for providing open data and other forms of access to their data. Companies successfully developed business models based on opening their data to the public sector. The company Uber provides a service called Uber Movement to share its data from billions of customer trips with city administrators. In Europe, five cities are participating so far –Amsterdam, London, Manchester, Paris, and West Midlands (UK) as of January 2019.[14] This arguably serves its interests as a way to gain favor from city administrators that often govern taxi and hired car industries. Private companies can also partner with research institutions to generate actionable insights concerning the development of new infrastructure-related business. For example, the company Scania A/B provided researchers access to a dataset of 855 million GPS pings emitted by nearly 60,000 vehicles for a study aimed to determine optimal locations for static and inroad vehicle charging systems in Stockholm (Shreenath & Meijer, 2016). In 2018, Google released datasets it currated included Open Images V4[15], a large set of images plus categories, human-generated labels, and bounding boxes; Atomic Visual Actions (AVA)[16], a large set of video annotations for labeling human actions and speech; HDR+ Burst Photography Dataset, and Google-Landmarks. They also aunched a beta Dataset Search tool[17] to help its developer community more quickly find large datasets useful for their projects. (Dean, 2019.) Sharing datasets and providing a tool for finding new datasets arguably serves Google's overall business interests by fuelling outside developers use of Google-provided tools to solve difficult tasks.

3.2 BIG DATA IN THE PUBLIC SECTOR

According to one interviewee, a driving motive for the public sector to develop new ways to derive insights from big data is to shorten the time it takes to create a usable insight or indicator for policy makers. Shortening indicator production time can provide decision makers with more up-to-date information regarding a given policy domain. Because statistical information is based on observations about the past, many indicators lag by months and sometimes years. Gaining a reliable view of the nearer-present supports the making of policies with appropriate impact and fit to the

[14] https://movement.uber.com/faqs?lang=hi-IN and https://medium.com/uber-movement/tagged/data-stories (Accessed 31 January 2019).

[15] http://ai.googleblog.com/2018/04/announcing-open-images-v4-and-eccv-2018.html (Accessed 31 January 2019).

[16] https://research.google.com/ava/ (Accessed 31 January 2019).

[17] https://toolbox.google.com/datasetsearch (Accessed 31 January 2019).

situation its possible futures. Such sped-up processes will entail public sector actors to develop their capacities and practices for utilising big data.

The ways datasets are produced and used in government are changing what is possible or expected in territorial governance (Ozga, 2009; 2012). As a policy instrument, data continues to grow in strength, speed and scope. For this reason, data-driven governance and policy-making is increasingly dependent on knowledge derived from data. Governing based on data, however, is not straightforward: datasets require constant attention and effort to be translated into actionable insights. New approaches to developing insights from big data are being explored through academic research and international pilots but in order for such developments to be brought into governance processes, public sector organisations will need data capacities. Utilising big data places new demands on public sector operations and resources.

Administrative bodies with fewer big data capacities and territories with fewer sources of data, due to – for example – fewer installed sensor arrays, slower network connections, or lower market penetration of mobile or IoT devices and network services among individuals, are at risk of missing out on future possibilities that are expected to become available as big data analysis evolves. This unevenness may lead to future discrepancies in the kinds of insights which can be produced by LAU, regional, national, or supra-national governing bodies. (Schintler, 2018, 340.) The digital divide between urban and rural areas, for example, could grow wider as these advanced data analysis systems and practices are adopted more rapidly and effectively by bigger, wealthier cities. Across EU member states, an unevenness of digital capacities could result in serious and structural development disparities. However, when governance is practiced from a cross-border, multi-administration, and multi-scale framework, these concerns of uneven development can be addressed by establishing meaningful settings in which large cities can share their data science capacities with small towns and disperse rural settings. Distributing big data capacities more evenly can enable public sector actors at multiple scales to make more effective impacts on future developments.

Good practices for big data in public policy can be broadly categorized into recommendations for organizational capacity and research design. In order to succeed in either of these categories, actors need to be aware of the limitations of big data and the ethical risks such as unintentionally harming certain populations through undetected algorithm bias and violating individual privacy (Crawford, 2013; Hargittai, 2015).

Following best practices for saturating organizational decision-making at all levels with access to high-quality data science can help overcome this gap. In the private sector, organizations with more than five years of experience in using AI-driven data analytics tend to hire specialized internal data science teams, task those with building machine learning models, establish checklists for new analytic models that include testing for transparency, privacy, bias, and give those teams autonomy in evaluating their success or failures (see Lorica & Nathan, 2018). For organizations starting from zero, developing internal data analytic capacity can benefit from focusing first on establishing the “first-mile” of organizing data acquisition and a programme of analytic projects and the “last-mile” of establishing an understanding of end-user needs and highest-value targets for application (Chui et al., 2018). (Organizational practices are discussed further in chapter 5.) For the purposes of policy-making for a growth corridor, these last-mile actors include all policymakers in various administrative bodies who can influence its development.

3.3 GOOD PRACTICES FOR USING BIG DATA IN CORRIDOR DEVELOPMENT

Big Data is widely discussed as a pathway for generating new and actionable insights. Yet, few private and public organisations have much experience using the advanced data analytic tools required to make such insights (cf. Lorica & Nathan, 2018); while processes for preparing raw data to make it usable can be long and require advanced data-related skills (Ahas et al., 2008)[18]. Meanwhile, there are some central challenges in utilising big data for corridor development related to, e.g., the missing spatial dimension from big data components (Shi et al., 2017) as well as to the fact that the majority of big data applications are designed for businesses and industries rather than governments. One of the reasons cross-cutting these challenges is that data management structures are too siloed in organisations and there is not enough institutional support for collecting, preparing, and sharing datasets (Giest, 2017).

In addition, some have argued after reviewing the situation that there is not enough big data suitable for territorial development or enough knowledge about available data (see Santala, 2016). This targeted analysis found that while there are now many municipalities, regional governments, and nations running open data portals across Europe, there are not many examples of raw from the source data streams or master data being made available in these systems. Instead, most datasets available on these portals are aggregated representations of data in the form of tables or spreadsheet reports made for purposes which are often too specific. Furthermore, datasets available through these portals may not be similar enough to each other to be combined for the purposes of cross-border territorial policy development. The task of combining datasets is also hampered by open portals having differing terminologies in their information architectures or categorization schemes. Initiatives to harmonise such categorization schemes across EU nations via initiatives such as Eurostat should continue.

Despite these challenges, big data is being used by researchers to generate insights in policy areas such as those related to transportation and infrastructure, smart mobility, economic growth, sustainable development, smart grid, energy efficiency, education and security (Kim et. al., 2014; Shi et al., 2017) – all of which are of key concern to corridor development. Examples of currently active European projects include SoBigData focused on strengthening the data ecosystem

18. According to Ahas et al. (2008), a research team needs to be able: ‘(a) to access the operator’s database; (b) to cope with the peculiarities of mobile operators’ hardware and software; (c) to work with huge databases; (d) to handle data security; (e) to be familiar with GIS, geographical data and statistics; (f) to handle social sciences methods; and (g) to address the needs of end-users (academic or applied).’

concerned with social data mining[19] and Coastal urban development through the lenses of resiliency (CUTLER) focused on implementing pilots for evidence-based policy-making in Thessaloniki, Greece; Antalya, Turkey; Antwerp, Belgium; and Cork County, Ireland.[20] In the U.S., the DSAPP at University of Chicago directly works with policymakers to address precisely defined problems such as predicting fire risk of buildings, predicting urban decay or blight, and reducing the number of 'water shutoffs' occurring in a municipal water system.[21] In the European context, policy concerning managing the balance between built-up human settlements and natural ecosystems has been supported using insights generated from Copernicus land-use data (Copernicus Land Monitoring Service, 2019).

Globally, academic researchers are finding innovative ways to model and address policy related problems through data analytics. The public sector can also take inspiration from the research designs of several research-oriented big data studies. Looking again to the policy contexts identified by the project stakeholders (see chapter 1.3), several such studies are notable. For land-use, a study data retrieved from space-based remote sensing to produce scenarios concerning how land in an urbanizing county in China could most efficiently be allocated (Chen, Sun & Sajjad, 2018). Along these lines, population and commuting statistics calculated based on passive mobile positioning data have used to defining functional areas, a study that served as an input to administrative reform informing Estonian Regional Development Strategy 2014-2020, the National Spatial Plan of Estonia 2030, and county spatial plans. For mobility infrastructure, a case in Sweden uses many GPS pings from a set of vehicles to help answer the question: 'Where should new in-road electric vehicle (EV) charging station systems be located within the corridor in order to maximize their use by customers?' (Shreenath & Meijer, 2016). In Estonia, research using mobile positioning data informed the renewal of the public transportation lines in Tartu. For economic growth, a recent Eurostat pilot project utilized detailed price data generated from price scanners (so called point-of-sale data) in Netherlands, Sweden, Norway and Switzerland to see how this data could support the Harmonized Index of Consumer Prices (HICP) (UN GWG Big Data, 2019). Cross-cutting many policy contexts, social media data has been combined with other forms of temporal-spatial data to produce insights into the availability of groceries over time by commuting mode, tourism preferences in nature preserves, and informing forest conservation planning (e.g. Järv et al., 2018; Hausmann et al., 2018; Lehtomäki et al., 2015).

Private-public cooperation also holds potential to generate new insights for corridor development for policy makers. Such experiments can introduce actors from both sectors to new ways for combining physical, social and digital dimensions of spatial development into data-enabled services. Out of many possible examples, two good practice cases were identified in the research area of this targeted analysis – MaaS Global & Whim App and Vainu (Box 3.2).

[19] http://project.sobigdata.eu/ (Accessed 6 February 2019).
[20] https://www.cutler-h2020.eu/pilots/ (Accessed 6 February 2019).
[21] https://dsapp.uchicago.edu/projects/ (Accessed 6 February 2019).

Box 3.1: Good Practice Examples

MaaS Global & Whim App: This company aims to reduce the need for privately owned cars by providing convenient door-to-door, multi-modal mobility service for city inhabitants and more consistent system usage for mobility providers (Milne & Watling 2018: 7, citing Ambrosino et al. 2016 and Heikkilä 2014). A notable practice of the company that it publishes an open API which integrates data streams from mobility service operators such as public transit (busses, metros, trains, trams etc.), bike sharing services, car sharing services, and taxis etc. for the markets it serves and offers clear instructions for how to use it. Another key practice is to partner with cities, regional governments, and public transit authorities to launch its service. The MaaS API motivates these mobility providers to share their data in a consistent and real-time way in order to be included in the Whim App service offering. In response to MaaS Global needs, Helsinki public transit system (HSL) launched an API any registered app developer can use to sell tickets and season passes. As of this report, the MaaS Global is operational in Helsinki, Amsterdam, Antwerp, and West Midlands (UK). They have ambitions to offer seamless door-to-door mobility services around the world. In October 2018, the company announced it will launch in Singapore.

Vainu & Uusimaa Region: The company Vainu sells a platform that businesses can use to identify quality leads, follow the activities of those leads, and predict their next actions based on data aggregated from a wide variety of data sources including company websites, news feeds, social media, and other records.[22] The platform visualizes data in a topically customized dashboard and interactive map. According to an interview, Uusimaa Regional Government (Finland) mentioned their organisation cooperated with Vainu to make an experimental interactive map showing which businesses and business sectors were growing or declining in which locations. This pilot project matched business IDs from the official business registry with the companies tracked in the Vainu platform. An objective was to predict growth and decline of sectors in the region. The regional government ultimately did not use the insights generated because the tracking data provided by Vainu was not comprehensive enough to cover all companies existing in the region. The pilot did however demonstrate a pathway for building new kinds of monitoring tools the public sector can use to track and forecast leading indicators on a geo-spatial and temporal basis. Especially when the private partner in such a cooperation is operating internationally and the public data used is comprehensive, such systems could cover the business activity of an entire growth corridor. The good practice to learn from this case is to try new private-public partnerships, identify the gaps in the outcome, and try again to grow the capacities of all actors involved.

The main point of these examples is that public organisations and private companies can work together to co-create new data-driven tools that provide new ways to generate policy-relevant insights into the interactions and flows occurring within growth corridors. These cooperations can lead to long-lasting and mutually beneficial outcomes. Even when an experiment doesn’t produce an optimal result, awareness of the potentials of such tools can foster developments and changes in practice that can ultimately produce value to corridors. A key to the success of such collaboarations is to aim toward producing new tools that support the policy-making needs as wide a set of actors within a growth corridor as is possible. In other words, developing new services in partnership with private companies should not be done to compete with neighboring regions or municipalities but rather to produce valuable new insights that benefit all regions and communities along the growth corridor.

[22] https://product.vainu.io/, (Accessed 23 January 2019).

3.4 PATHWAYS FOR ANALYZING BIG DATA TO INFORM INTEGRATED POLICY-MAKING IN GROWTH CORRIDORS

Returning to this project's conceptual framework (see chapter 2) and key themes identified in the stakeholders' needs analysis, a categorisation of big data sources for corridor development is proposed (Table 3.2). In this scheme, corridor interactions and flows are assigned to physical, social and digital dimensions to which they mainly belong. However, assignment to one dimension does not necessarily mean a dataset cannot be used to produce insights in the others. For example, social media data can fit any three of the dimensions depending on which parts of the data are emphasised – e.g. a Facebook social graph (social) vs. Tweets over time (digital) vs. the contents and geotags of Instagram photos (physical).

Table 3.2: Categorization of example data sources for corridor development

Dimension	Interactions/Flows	Example data sources
Physical	Cargo Flows	Traffic management systems, maritime shipping manifests, air cargo manifests, waste and recycling metrics
	Customer Flows	Retailer bonus programs, price scanner data, space-based imaging of retail parking lots
	Daily Commuting	Vehicle-produced data (e.g. from engine computers or self-driving cars sensors), mobile positioning data, private GPS services, electricity consumption by building and unit
	Student Mobility	Student train/bus ticket sales, university student registers, exchange programs
	Migration & Relocation	Postal service records, population registers, data about issued residence permits, home purchases, rental listings, mobile positioning data
	Tourism & Leisure Flows	OECD tourism statistics by nation, Eurostat tourism data[23], national border crossing statistics, hotel night stays, device languages on wireless networks, social media photos of tourist sights, geotagged tweets, mobile positioning data
	Business Travel Flows	EU border crossing statistics, tradeshow 'booth lists', customer data from booking agencies
Social	Research Cooperation	Lists of funded research projects, research consortiums, bibliographic datasets (e.g. co-authors & institutions), open science portals, patent data
	Business Creation Interaction	New records in national business registers, llists of startup contest attendees, listed startups hosted at incubators, startup crowdfunding sites, feeds from international startup blogs (e.g. ycombinator) and regional startup blogs
	Trade Interaction	Trade statistics, B2B transactions, industry conference lists, tradeshow listings of registered booths, speakers, or attendees

[23] https://ec.europa.eu/eurostat/web/tourism/data/main-tables

	Inter-Firm Cooperation	Supply chains, business consortiums, shared patents
	Family/Social Interactions	Population register, mobile positioning data, social graphs (e.g. Facebook Social Graph, LinkedIn network connections, Twitter followers)
	Investor to Business Interaction	Stock market data, crowd-investing platforms, AngelList, Crunchbase, investor call transcripts, lists of startup investors,
	Student Interactions	Course lists, Student union member lists, exchange student lists; joint degree programs by two or more higher eduation institute, student groups on social media, registers of student organisations
	Planning Cooperation	Minutes of official planning meetings, verbatim transcripts of public meetings, vendor bids, architectural renderings, models
	Purchase flows	B2B invoices, price scanner data, import/export metrics, personal finance/fintech apps
Digital	Network traffic flows	IP Pairing Site Network Traffic (IPX), Wireless Data Traffic; Internet Service Provider networks traffic; Satellite signals
	Monetary flows	IPO trading, corporate bonds, public bonds, business loans, consumer loans
	Communication flows	Phone utility metadata, mobile positioning data (e.g. caller-called pairs), top web searches (e.g. Google Trends), user interactions with social media posts (e.g. Instagram, Twitter, Facebook, Mastadon), languages of social media posts, top website rankings by region (e.g. Alexa), mobile device languages, multilingual news and information (e.g. EC JRC Europe Media Monitor)

This categorisation, by emphasizing flows and interactions, bridges from the conceptual framework to policy themes that are highly relevant for growth corridors: 1) transportation and infrastructure planning; 2) regional economic development, and; 3) land-use planning – the major corridor development dimensions identified by Zonneveld and Trip (2003). Fitting the aims of this targeted analysis, the categorisation is also intended to help identify new examples of datasets that could describe the functionalities of corridors and provide new insights for data-driven policy-making. Again, the newness of a dataset depends greatly on an organisation's perspective which is shaped by the data it currently uses, the combining datasets that have not been combined before (e.g. via temporal data), and the accumulated depth of an old and well-maintained dataset. While several examples are given, the categorisation can also be used as a starting point to imagine, speculate and confirm whether other datasets exist that can describe the interactions and flows. In some cases, hopes for datasets could be made to exist by bringing a relevant set of actors together.

While this categorization supports wider thinking in selecting datasets for producing insights, it does not immediately answer questions about how useable any given dataset would be for a policymaker. To make that kind of assessment, a multi-factor review is needed based on an abstract-level typology of big data characteristics. To this end, this targeted analysis proposes the typology of data characteristics described in the following subchapter.

3.4.1 Characteristics and typology of big data

Today, there are many kinds of datasets and many approaches to categorizing them as is evidenced by the already mentioned differing information architectures of dataset portals. This project proposes a typology of key characteristics of big data datasets (Table 3.1).

Table 3.1: Typology of new data sources by variables and ranges of attributes.

Data Categorization Variables	**Range of Attributes**
Availability	Open data ↔ Available only under special contract ↔ Purchasable proprietary data ↔ Unavailable proprietary data
Comprehensiveness	The Data Sample is Limited (only some subset of data subjects or entities of all possible ones is included) ↔ Dataset is Complete (includes all possible data subjects or entities, e.g. a population register)
Level of Processing	Raw (e.g. direct from sensors) ↔ Pre-Processed Data ↔ Processed Data ↔ Highly Processed Data
Intended Audience	Machines ← Programmers of Machines → People
Observational Qualities	Direct Observation ↔ Synthetic
Level-of Detail	Fine-Grained (e.g. vehicle journeys) ↔ Rolled-up (e.g. average household income by zip code)
Level of Structure	Highly Structured ↔ Semi-Structured ↔ Unstructured
Refresh Frequency	Instant ↔ Milliseconds ↔ Daily ↔ Weekly ↔ Monthly ↔ Quarterly ↔ Annually ↔ Every few years
Confidence in Updates	Low Confidence (updates are uncertain) ↔ High Confidence (updates are highly likely to occur)
Extraction Effort	Great effort is required to extract data ↔ Little effort is required
Analytical Effort	Great effort is required to analyse data ↔ Little effort is required
Clarity of Ownership	Ownership is clear and singular ↔ Ownership is clear but shared ↔ Ownership is clear on paper and unclear in practice (e.g. "Do I or Facebook or 3rd party app makers own my Facebook data?") ↔ Ownership is unclear or no longer traceable
Spatial Resolution	Exact location, Local, Neighbourhood, Municipal, Regional, National, Multi-national, Continental, Global
Temporal Resolution	Milliseconds ← … → Decades

Box 3.2: Explanation of variables in the typology of new data sources (Table 3.1)

Availability ranges from freely available open data (e.g. data published on an 'open data' portal) to purchasable proprietary data owned by a company to unavailable proprietary data. To some level, demand can influence how available a dataset is and under what terms.

Comprehensiveness refers to how completely a dataset covers the totality of data subjects or entities within the scope of an inquiry. For instance, Tweets made in a given time frame and spatial dimension will only include statements made by Twitter in during that time and in that space, while statements made by other people through other channels are excluded.

Level of Processing is a spectrum from rawness to highly processed. Data streamed directly from a set of sensors is an example of raw data while data that has been evaluated and modified by data scientists in order to ensure its veracity and prepare it for use is an example of highly processed data.

Observational Quality of the data is a spectrum going from directly observed to completely synthetic data. Synthetic Data is derived from some original aggregated data source to approximate finer grained details. Examples include taking average income data for a neighbourhood and distributing it as average incomes for every building in the neighbourhood or taking rough grain data about a vehicle's travel path and velocity and approximating locations between recorded samples. (Grinberger & Felsenstein 2018, 109; Hwang et al. 2018, 135.) In a similar fashion some datasets are used as a proxy for other datasets.

Intended Audience ranges from data designed for machine use to data designed for humans to read. For example, data can be generated by one machine for another machine and transmitted in a form not readable by humans (e.g. a self-driving vehicle interacting with a Traffic Management System) or data can be input into a system by a human for another human to read (e.g. social media posts, email messages, online survey answers, or chat messages). In the middle are data structured in a format that is both legible to people skilled in ICT development such as RSS, JSON, XML or other formats and data structured in reports ranging from spreadsheet documents to PDF reports.

Level of Structure ranges from highly structured with consistent fields and value types for every row of data to semi-structured with inconsistencies between rows in how data is recorded to completely unstructured with fields lacking any form of delimitation. Structure can also refer to the type of database ranging from relational meaning the data is stored in a set of tables linked together by a primary key field (e.g. SQL) to non-relational (e.g. noSQL). Another aspect of structure can refer to the completeness of metadata describing details of what each field contains.

Refresh Frequency ranges from dynamic data that is updated in near real-time to static data that is updated far less often. Dynamic data is more necessary, e.g., for highly interactive contexts or studies that aim to deliver on insights about 'now', the recent past, or forecast the near future. Static Data is sufficient for other kinds of study.

Confidence in Updates is a user's perception of how likely the originator of the dataset will follow through on updating it in the future. As a form of anticipation, a user will usually base their confidence level on evidence such as the reputation of the organization producing the data, whether the dataset has been updated before, and an assessment of how frequently other datasets appearing in the same portal or system have been updated.

Extraction Effort refers to how much focused energy is required to extract the data from a given source (e.g. unstructured data in printed material).

Analysis Effort refers to how much staff time, computing power, and technical skills are required to analyse the dataset. It also includes the time it takes for pre-processing datasets or building applicable models.

Clarity of Ownership refers to how obvious it is what entity owns the data. This seemingly straightforward concept becomes blurred by user perceptions about their own data. For example, a user of a social media site may hold a belief that they own their posts and photos. However the terms and conditions for using the service may say otherwise, and the service provider may have the right to use or sell their data based on legal contract. Ownership of data also impacts long-term reliability of live data, as live datastreams maintained by private companies can disappear if those companies happen to go bankrupt or simply decide to stop supporting the stream (Erhardt et al. 2018).

The **Spatial** and **Temporal resolutions** of a dataset refer to how fine-grain or course the data is (e.g. 1 km grid vs. 72 cm grid) and how frequently observations are recorded (e.g. twice a day vs. every milliseconds). These resolutions impact how detailed or granular an insight can be generated and in practice often need to be adjusted in preprocessing when combining datasets of different resolutions.

The variables of the typology (described in Box 3.1) are generated from a literature review covering many articles concerning cases of big data applied to regional development, urban planning, and policy development (e.g. Shintler & Chen, 2018). These variables and attributes can help policy-makers and their data teams assess the suitability of datasets for generating a specific kinds of insights and for identifying approaches for combining them.

Depending on what kinds of new insights would be most beneficial to a given policy context and set of policy questions, some attributes for these variables make some datasets more desirable than others. For example, if a policy study requires historical depth at a coarse grain temporality, a temporal resolution of milliseconds is less useful than an annual resolution. Whereas, a much finer resolution is required to make so called nowcasts – or forecasts of near future developments. Taking the datasets used later in this report for the case studies plus a few new datasets found in literature, Appendix 2 was produced to demonstrate how some set of the variables from the typology can be used to assess the value and viability datasets for this project's research area.

This project proposed tools for assessing the value of new datasets for policy-making. Data categorization framework sets the stage for public sector data science which can surface hidden phenomena occurring in growth corridors. Currently, the evidence-based decision-making relying mostly on data about mobility of people and goods has directed attention to the development of corridors as platforms for physical flows – i.e. transportation corridors – and the enhancement of physical movement, instead of exploring the factors that explain the causes and motivations for physical movement (Kalliomäki & Forsell, 2017). At the level of European growth corridors this approach is, on the one hand, well justified due to the European level attempts to ease transnational mobility. On the other hand, however, new datasets related to social and digital interactions capable of explaining the factors behind physical movement might contribute to the development of more environmentally, socially and economically sustainable territorial practices in corridors functioning as platforms for numerous interactions.

From this perspective, the factors causing the movement at the growth corridors should be at the very core of corridor development due to the fact that these issues cannot be affected by the individual actors at the lower scales of territorial development. Overall, the comprehensive approach to spatial development, and the manifold questions related to the utilisation of big data in territorial policy development in the European growth corridors highlight the importance of policy integration in corridor-based development processes.

4 CASE STUDIES EXAMINING THE POTENTIALS OF NEW DATA SOURCES

4.1 CASE 1: TRAFFIC MEASUREMENT DATA IN FINLAND

Box 4.1: Summary of Case 1

Objective: To examine the potentials of automated traffic measurement data for corridor development by analysing the origin-source traffic profile of E18. The framework is intended to enable a later coupling of traffic data with economic parameters of districts. In the future, it could also be combined with data depicting wider corridor dynamics. The ultimate goal is to understand how and when people and goods move, at daily and annual resolutions.

Data used: Traffic intensity data detected by 79 below-the-surface induction loop sensors between years 2010 and 2017 along the E18 route from Turku to Kotka, Finland. The data is provided by the Finnish Transport and Communications Agency (Traficom). The data is rich in details and spans several years. The data has many good attributes that makes it interesting for further research.

Novelty of the approach: The proposed method estimates the Origin-Source (O-D) matrix indirectly from vehicle passage information. The accuracy of the results can be estimated based on the spatial configuration of the induction loop sensor network. Existing methods require more specific data, e.g. traffic polls and full coverage of all road junctions by either traffic sensors or observers.

Challenges: Because the data cannot distinguish individual trips, but only anonymous passes, validation of the results based on direct observation is not possible from the data alone. Further research is needed for the scientific validation of model used to produce the O-D matrix.

Implications for policy-making: The availability, reliability, and depth of the data makes it suitable for policy-guiding tools and analytics. The network of stations generating data is still sparse on some roads, but for E18 and the NGZ this system is well established and enables analytical tools to be applied to gain more insight to policy-making. Predictive models are among the toolset, that policy makers can utilize with researchers to leverage this dataset for many applications. The measurement networks is also stationary, so the data can be used in tandem with other datasets that describe the growth corridor dynamics. These types of indicators are for instance local economic factors. The proposed time-space pixel format is one among many possible means to pan-European traffic analysis. This would help in infrastructure planning (a further stage) for corridor development. Especially the trend analysis over the O-D indicators is recommended.

4.1.1 Background and objectives

The objective of case 1 is to examine the potentials of automated traffic measurement data for corridor development by analysing it to produce an origin-source traffic profile of highway E18 in Finland. While transport and traffic information has been typically collected through surveys, in the future, the volume of other forms of traffic data will increase significantly due to an increasing number of internet connected devices, vehicles and sensors installed on mobility infrastructure. Combining data from these sensors will make it possible to measure the functionality of road network with greater precision.

For example, cargo-related sensors allow freight traffic to be measured and monitored at more precise scales. Nowadays statistics on traffic and freight rates are almost exclusively based on the total volume of transport, which makes it very difficult to make more detailed analyses, for example, of interregional interaction and functionality. However much of this data is unavailable because it is collected by private firms that do not share it with cities and public actors.

There are many existing methodologies for analysing traffic flow to make an O-D matrix. This study follows Parnami (2018) which uses deep learning (DL) to analyse a traffic flow time-series from a selected subset of spots in a city-wide traffic network. There are known challenges in the proper multi-scale analysis methods of traffic flows and temporal data which are also described in extant literature. For instance, Boyandin et al. (2011) point out that many techniques for the exploration of spatio-temporal data have been developed, but they prove to be only of limited use when applied to temporal origin-destination datasets.

The traffic analysis is considerably more varied than just about producing O-D matrix estimates. Basic approaches can be grouped to at least following categories: time-series analysis (e.g. Tang, 2014); radial basis functions (e.g. Wang, 2003); trend analysis (e.g. Li, 2015); Statistical analysis (e.g. Bera, 2011); general stochastic analysis (e.g. Chen, 2014); minimum description length (MDL), and; traffic simulation (e.g. Vissim, 2017). Most of the existing traffic analysis focuses on the dynamics of a single point (a junction etc.), or drawing a correlation between two locations, or analysing a limited area in depth over a chosen period of time. The methods vary from survey polls and video surveillance to traffic simulation. The latter two require a rather high level of road network detail. The underground traffic sensors are a problematic information source, since they are seldom placed on the junction areas but a bit further away. There exist many analysis tools for communal traffic, but only few (if any) could be applied to the European level. Our research omits the traffic congestion problem, since several aerial, video surveillance and sensor solutions already, exists for that.

We are focusing to produce a generic multi-scale view over the traffic flow, which can allow the extraction of dynamical dependencies between districts at a further stage. An important choice is whether to record the traffic fluctuations (especially the speed) or whether to try to estimate the origins and destinations of the traffic. This study focuses on the latter.

The dynamic origin-destination (O-D) matrix consists of a matrix, where individual entries are pairs of districts and values tell the activity of commotion. Having this kind of O-D available for example to trend and correlation analysis would benefit the district and regional planning greatly. Achieving an O-D matrix requires either direct methods (domestic and roadside polls and license plate observation) or indirect methods. Our objective is about indirect O-D estimation based on traffic sensors, which are induction loops placed beneath the road surface. The most common indirect methods are statistical, and they are listed in Bera (2011).

4.1.2 Materials

The materials of the case study consisted of the Automated Traffic intensity Measurement (ATM) data provided by the Finnish Transport and Communications Agency (Traficom). Similar ATM data is gathered e.g. by Swedish Transport Administration (Trafikverket) and Estonian Road Administration (Maanteeamet). The data is produced by induction loop units installed underneath the road surface of the E18 route. The Digitraffic interface in Finland provides open access and up-to-date traffic information on the Finnish roads and railways. The following traffic and condition data are available through the Digitraffic interface: up-to-date fluency information; history of the previous day's fluency; the average fluency data of the previous 12 weeks; up-to-date ATM data; up-to-date free speeds; up-to-date measurement data of road weather stations; status information of road weather stations, weather forecasts for road sections, and; disruption bulletins. Further attributes of the data are detailed in Table 4.1.

Although traffic sensors were used based on induction, a similar analysis could be based on roadside magnetometers or a traffic camera system augmented with an automated vehicle registration, classification, and with tuning to individual traffic crossing lanes and details. Induction loop systems are of an install-and-forget kind. The hardware, installation and maintenance costs approximately equal within 10 k€ accuracy[24], but the installation of the induction loop interrupts the normal traffic. The actual costs analysis of different technologies has been omitted in this report. Spatio-temporal traffic flow data is more readily available today than before in regional as well as local decision-making processes and various project settings. Researchers can harness more powerful machine learning and AI tools and solutions to read and analyse even real-time big datasets. As Ren (2018) summarizes, these kinds of new approaches provide a way to discover the effects of atypical circumstances, such as road works and events. Moreover, the analysis results can be further used in cities' traffic planning and land use planning (Ren 2018, 263).

Table 4.1. Dataset typology applied to key dataset used in Case 1

Dataset: Traffic Management System Data from Traficom	
Variable	**Attribute**
Availability	Open Data
Comprehensiveness	Nearly complete: All vehicles crossing over sensors are recorded except for in rare cases in which a sensor is malfunctioning.
Level of Processing	Raw
Intended Audience	Machines
Observational Qualities	Direct Observation
Level-of Detail	Fine-Grained
Level of Structure	Highly Structured
Refresh Frequency	Daily
Confidence in Updates	High Confidence
Extraction Effort	Requires little effort to extract data

[24] US Dept Transportation (2018)

Clarity of Ownership	Ownership is clear
Size	20 Bytes / (vehicle passage) or 160 kB / (year x km) or 0.5 TB (total E18)

The traffic data used in this case was openly available for Finland only. However, researching other European traffic agencies it became clear that these measurement systems were embedded in other countries as well, thus indicating potential for transnational corridor analysis. This case study only summarized Finnish and Swedish data publishing policies by the responsible authorities. Briefly put, at the Swedish Transport Administration (Trafikverket) a database called STRESS is a system that only delivers real-time data to other internal systems. Finnish Transport Infrastructure Agency (Väylä) has a similar real-time service, but it also provides the open history view to the accumulated data via Digitraffic and Digiroad APIs. In Sweden, Trafikverket shares some interfaces from where the data leads to systems such as Datex and other APIs that publish it as open traffic data. Authorities however do not have a standard open data interface to access the history data. The Swedish Transport Administration gathers data from various ongoing research projects (e.g. Mobile Millennium, DriveMe and BADA).

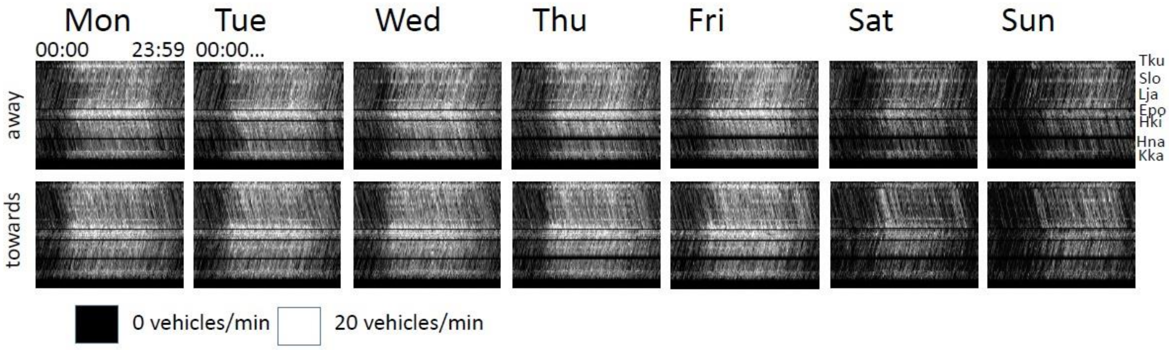
4.1.3 Methodology

The initial data has every vehicle passage recorded. The passage record holds the vehicle type, speed and direction. The vehicle type is approximated from the dynamic event observed by the induction loop. The data pre-processing is rather involved, as is usual in the ML projects with the dynamical data.

The first step is to transform the data to such a format, which allows easy visual verification and efficient computation. Marbán (2009) developed the cross-industry standard for data mining (CRISP). The data visualization is for detecting anomalies like road works but also for early data validation and enabling fast computations. A .jpg image format with 2 km and 6 minutes pixel size with the traffic intensity as the pixel intensity was chosen. The format is efficient both computationally and in terms of memory usage. Information from individual vehicle passages is lost, but the overall analysis is easier to make. A daily snapshot of the traffic over the whole length of the E18 road consists of two images (personnel vehicles and the cargo traffic). A temporal-spatial coordinate transform is then applied, which makes a typical vehicle travelling at the average speed of the traffic to appear as a straight vertical line. A multi-scale predictive model with spatio-temporal flexibility is then fitted to the images in order to simplify the clustering over different time periods (from 2 mins to years) and different geographic scales (from 2 km to 400 km in this case). The current base granularity was chosen for the analysis, and a more fine-grained approach will be tested in the future. There are relatively few truly multi-scale methods existing, one of the most influential is Liu (2003).

The traffic signal has fluctuations (noise), which seem to move approximately at the same speed as the main traffic. The traffic intensity fluctuates naturally, but there are also so called jamitons (selfstanding concentration waves, Flynn & Kasimov, 2009), which generally have a little slower speed than the main traffic. E18 road seems to have rather few jamitons, and we chose an approach, which is based on the observed speed of the normal traffic fluctuations. Thus, the traffic concentrations can be seen as light "threads" in Figure 4.1, which depicts both the traffic away and towards Helsinki on 23 January 2017. Steady black lines are from traffic sensors temporarily offline. The black lines are filled in for the later analysis by a simple linear interpolation scheme. The vertical axis is distance, from Hamina (Hna) to Turku (Tku), and the horizontal is time, from midnight to midnight. Altogether both natural variations and jamitons are somewhat detectable in the overall traffic flow image although the type b) are best detected using the traffic speed information. In general, the traffic data contains practically all of the traffic events (speed and vehicle type on each passage) and has potential for future studies.

Figure 4.1: Difference between weekdays in the traffic flow pattern along the route E18[25].



The chosen methodological approach has three steps:

1. **Traffic flow rectification**: The traffic flow (through road length and time) has an angle, which is dictated by the dominant traffic speed. An initial transformation rectifies the flow lines (see Fig. 4.1) so that the image noise is maximal in the horizontal direction and minimal in the vertical direction.
2. **Approximating a predictive model**, which includes the origin-destination structure in itself. There are two methods to create such a predictive model; a least-square regression using a set of traffic intensity rectangles (RRM), and a minimum description length (MDL) approach. The former resembles the GLS of Bera (2011) and the latter the IM (van Zuylen 1978)
3. **Constructing the O-D matrix** from the resulting model. The models (and resulting O-Ds) are different for the 5 business days, for the weekends and for various seasons. Separate models can be specified for individual districts and times of day.

[25] Difference between weekdays in the traffic flow pattern along the route E18. The week Jan 23th-29th on 2017 is in question. The upper row is the traffic intensity away from Helsinki, and the lower one towards Helsinki. Vertical axis is distance (380 km) and the horizontal axis is time (24 h). There were 79 TMS points along the 380 km road length of the route from Turku to Kotka.

The model is geographically generic, meaning that it is applicable anywhere where corresponding traffic data and similar economic indicators are available. Due to the constrained schedule no road network analysis (Omatu & Seinfeld, 1989) is covered but only the E18 road is included. Each measurement point has been associated to a small set of nearby municipal or postal areas in order to assign the economic descriptors properly in the future studies. At the moment the model suits to unsupervised learning enabling clustering of different municipalities and time periods. The supervised learning will commence when economic indicators are organised and available. Future tests with some economic data will further enhance our understanding of the full potential of the model.

4.1.4 Results

A novel method was applied to observe the traffic flows in the origin-destination matrix. As shown in Fig. 4.1, this was done in flexible spatio-temporal scale which is able to work with the data which does not yield to the O-D analysis provided by existing models. The O-D matrix (given proper data) is an excellent tool for analysing dependencies, similarities, synchronization and differences between districts, which helps the regional decision-making related e.g. to economic growth.

The results section has four topics: the accuracy of the O-D matrix, an O-D matrix example, a comparison of personnel and cargo intensities, and potentials for a wider application.

Traffic prediction error

The traffic prediction error over the E18 (380 km) and over 24 h varies between 2.7 and 4.0 vehicles/min for a period of one hour, depending on the method and the prediction time range (see details from the Appendix). This is a rather good accuracy, but inadequate for the quiet periods (night, midday). The accuracy is better when the analysis covers longer periods of time (same day during a month, during a season, over a year).

An O-D example between Turku-Helsinki

An example of the O-D snapshot over Monday mornings and evenings from Jan-Apr 2017 for the estimated traffic between pairs of districts. The origin is on the left column and the destinations are in the top row (Table 4.2). The traffic within each district (origin and destination at the same district) is possible, but eliminated from this result. This is because some districts have only one traffic sensor, making the O-D figure undefined.

Table 4.2: An Origin-Destination table for the west half of Highway E18[26].

	Morning 7:30-9:30 (vehicle/min)					Evening 16-18:00 (vehicle/min)				
O\D	**Tku**	**Slo**	**Lja**	**Epo**	**Hki**	**Tku**	**Slo**	**Lja**	**Epo**	**Hki**
Tku		5 1	1 1	15 2	20 2		9 1	0 0	4 0	8 1
Slo	10 1		1 1	5 1	10 1	6 0		2 0	3 0	5 0
Lja	1 1	3 0		10 2	10 1	0 0	1 0		5 1	5 1
Epo	3 0	2 0	3 0		60 4	0 0	0 0	1 1		2 0
Hki	5 1	4 1	5 1	50 3		2 0	0 0	0 0	3 1	

Each district is considered to be a 10 km (Slo,Lja,Epo) or 20 km (Tku,Hki) segment on the road. The O-D matrix produced by the method is based on the MDL theory, and in reality, there are several possible arrangements of the sensor set and road network, where the results would be erroneous. Thus one has to remember that the O-D matrix values are not actual physical traffic flows, but estimations based on the MDL principle. Their possible utility lies in a further trend analysis.

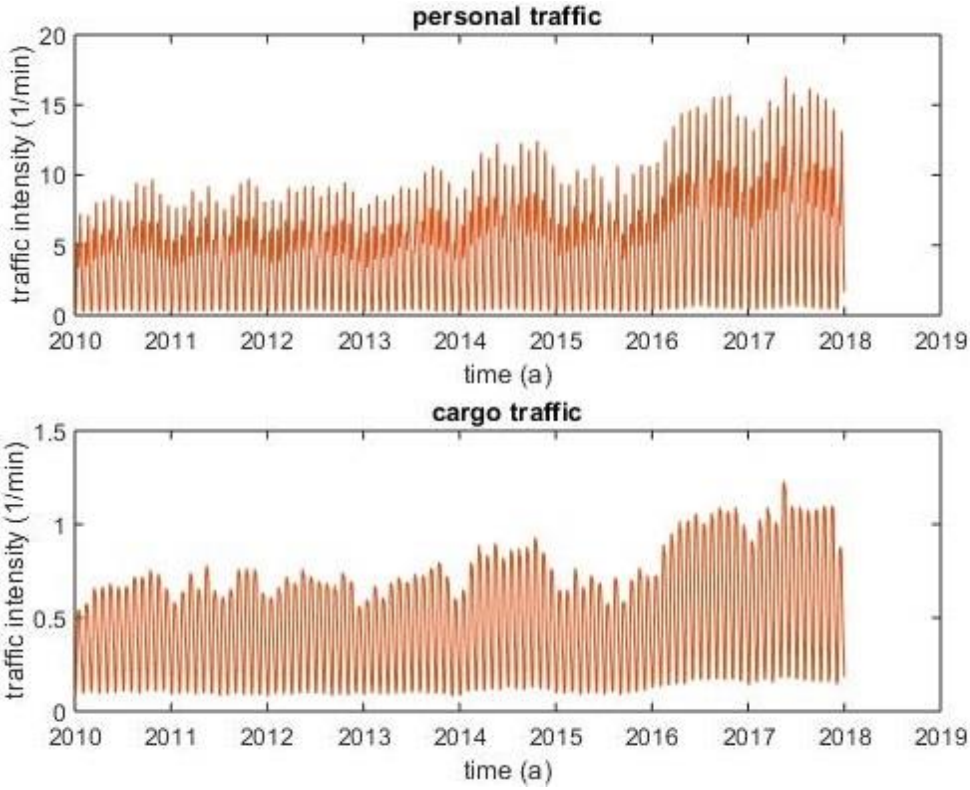
The accuracy of the results (again based on the MDL principle) is approx. ≤ 5 vehicles/min (personal traffic) and 4 vehicles/min (cargo). The cargo traffic O-D would require considerably longer periods of time to increase the accuracy of the O-D matrix. The O-D matrix is dynamic in the sense that the associated time window can be e.g. last 4 months. Furthermore, a measure of the O-D matrix is its daily symmetry. Most of the personal traffic seems to be daily pendulum commuting, meaning that the same vehicles return in the evening where they originated from in the morning.

Personnel and cargo intensities

Figure 4.2 depicts the frequency of the personnel vehicles (buses, personal cars) and cargo vehicles (trucks and lorries of all kinds). The total traffic flow is an average over the route E18 between Helsinki and Turku.

[26] An Origin-Destination table for the west half of Highway E18 on Mondays, January-April 2017 during the mornings and evenings. The districts Turku-Salo-Lohja-Espoo-Helsinki are presented as Tku-Slo-Lja-Epo-Hki. The black numbers are personal vehicles and grey numbers are cargo traffic.

Figure 4.2. Personnel and cargo traffic on the route E18 during 2010-2017. Each month have the average daily distribution presented.

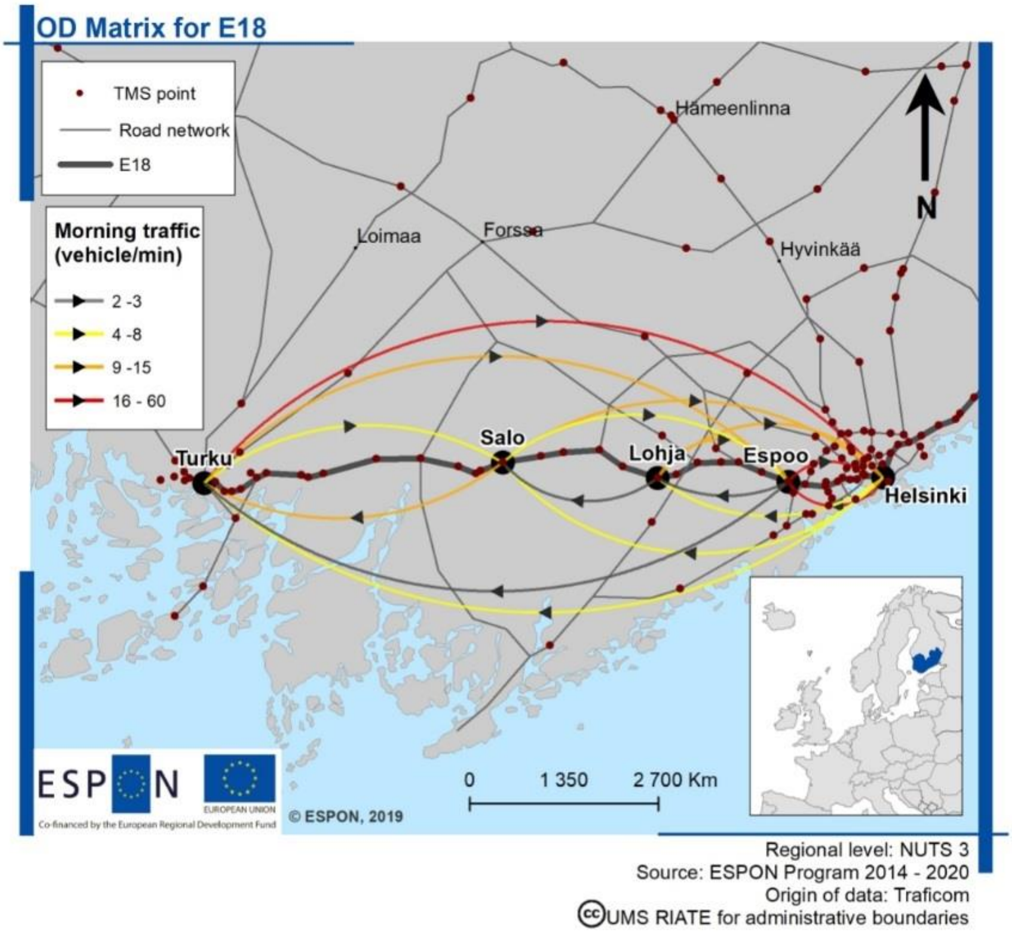


The years 2010–2015 are surprisingly monotonic, as seen from the Fig. 4.2. The years are rather similar under many other metrics, too. The beginning of the year 2016 shows 50% growth in the private traffic. Interestingly, the driving at night (between 02–04 AM), had 300% growth (not depicted in the Fig. 4.2), and this growth occurs mostly near biggest cities (Helsinki and Turku). This might be at least partially explained by traffic re-arrangement at the Western exit of the city in Helsinki. In addition, each year has a quiet holiday season reducing the traffic approx. 20%. This effect starts already on November. However, tracing the actual sociometric and economic causes for these changes was beyond this study, but overall the detected phenomenon indicate the usefulness of the detailed traffic data.

4.1.5 Potentials and limitations

Our current understanding is that expanding the analysis from the one single road to a network of roads (see Map. 4.1) will increase the O-D accuracy. If all network edges were equally traffic intensive, there would be only a little theoretical improvement, but since major roads dominate, one would get an improved estimate (and current or more inferior level of accuracy at the small roads). There are not enough traffic sensors to measure the flows at all the grey side roads, but in a generic application one should be able to utilize both the traffic cameras and the existing sensors.

Map 4.1: The OD Matrix for E18.



The case 1 is limited only to one route (E18). Usually traffic analysis concerns an interesting subpart (Deng & Cheng, 2013) of the total network. In the presence of the network, there are several possible (and sometimes rather equal) routes from one origin to one destination. Then the O-D analysis is not a matrix based, but it usually aims to network capacity maximization (Ahuja et al. 1993).

No direct validation of the results is possible due to the nature of the data, and thus further research is needed for the scientific validation of the O-D matrix. The validation of the resulting O-D matrix requires:

- either a field campaign with human observers,
- addition of traffic cameras with coverage and analysis software to record traffic flows in considerable detail,
- or some census material e.g. a cross-relation between the housing and work districts.
- Also, a dynamics comparison between economic data and traffic data is needed.

4.1.6 Implications for policy-making

The analysis is relevant for local and regional decision-making because the O-D information is of low-noise, with an accuracy estimate, and it can be generated in a chosen granularity level. The granularity here means both the spatial (from 1 km units to only main districts identified) and the temporal resolution (e.g. an O-D snapshot per every 2 hours to annual ones). On the NGZ one can see a lack of the micro-economical and micro-social data, and an integrating holistic large scale view over the both. One of the mismanaged resources is the traffic data, which could be a supportive signal in the process of explaining the micro-economical changes.

The following actions are recommended for the Finnish study area:

- Turku-Tampere, Karjaa-Salo and Karjaa-Turku traffic detectors are missing and should be implemented. The detectors can be of roadside or in-road variety. The added detectors would improve the traffic image and help understand the local economical couplings. The E18 road is not enough for that purpose.
- Existing data from Helsinki-Lahti-Kouvola and Helsinki-Tampere should be gathered for the analysis similar to this case 1, to provide a more complete picture over the Southern Finland.
- The long distance public transport data should be gathered and provided from e.g. Helsinki-Tampere-Turku-Tampere and Lohja-Karjaa-Salo routes. The data should be in the O-D format. This additional data source would help in comparing especially the total trafficking person count.
- Integration of the mobile data, population income profiles and the traffic data to a publicly accessible form. This would be partially a research effort involving universities, but also a technical matter involving the data providers and some stakeholders and experts. The mobile data quite likely offers similar but geographically more unified view to the daily pendulum traffic and economical and cultural couplings between various districts.

In addition, applying an integrated mobile and traffic data analysis leading to localized O-D summaries on the European level is recommended. This would provide a useful review on various infrastructure projects and to the political decision-making processes. However, an integrated mobile and traffic data analysis would require standardization and implementation of efficient and utilization-friendly national interfaces in several countries. The traffic data is only one example of detailed temporal-spatial data. As opportunities related to open data policies become known to various governmental institutions and private companies, we may see a shift toward more open traffic datasets in Europe. In this context, preliminary methodological development conducted in the case 1 is a step towards more comprehensive utilisation of these datasets.

The proposed analysis method of the traffic measurement data could provide spatially detailed trend analysis. Trend analysis can then be used in a traditional way to help in transportation planning and improve the models used in the existing traffic analysis tools. Mapping the bottlenecks in traffic routes is the most direct application of the traffic flow data itself and does not require the O-D tool. There is a need for more research to extend the analysis to the traffic networks. In addition, communication between policy-makers and researchers is needed to discover and apply best practices to use the model.

After an increase of the sensor network coverage, and augmenting the sensor network with magnetometers, drone surveillance and ordinary traffic cameras, one would increase the sampling coverage of the road networks and therefore the O-D matrix accuracy and coverage to a level, where very fine-grained regional and district level analysis would be possible. The main value of the O-D matrix would be realized when it would be coupled to economic models of the districts in question. This would require a further integration of the publicly available data. Also, the localized economic data such as general census data and GDP need to be available as open data. The temporal frequency of public economic data produced is often too sparse. A variety of frequently updated economic indicators are needed before the full power of traffic models can be harnessed.

At the moment traffic data, localized economical models and population statistics seem to come from national interfaces (if any). A truly European modelling and analysis of growth corridor dynamics requires an enormous data preparation overhead. This stems from the reason that these datasets are still generated, managed, utilized and processed by various operators. The central EU administration should enforce policies, which make the above mentioned three data sources an easy commodity to the research community.

4.2 CASE 2: PROJECT PARTNERSHIPS IN THE EU

Box 4.2: Summary of Case 2

Objective: The core objective of this case study is to conduct experiments with European Regional Development Funding (ERDF) data to analyse the interactions, volume and balance between the partners and to evaluate the usefulness of this data for corridor development.

Data used: Aggregated project and beneficiary data was retrieved from the data portal called keep.eu jointly operated by the Interreg program and the Interact Harmonized Implementation Tools (HIT). The project data acquired from programming period 2014-2020 contained data about 2,353 projects with a corresponding 18,318 partners across the EU.

Novelty of the approach: The social network analysis approach widens the understanding about spatial interactions and connectivities between regions, which is often lacking from corridor development activities.

Limitations: This case study was conducted partly with the whole dataset and partly with certain selected ERDF thematic categories that were utilised to analyse project partnerships in corridor specific thematics. However, there is potential to extend this analysis approach with machine learning and richer semantic representations of various development projects, as well as to combine, e.g., diverse funding instruments to the analysis, e.g., through Cordis. In addition, factors explaining the geography of collaboration should be further studied.

Policy implications: The experiments conducted in this case study demonstrate how big data analysis can reveal the spatial dimension of partnership networks which can support integrated policy-making that fosters greater corridor development and cohesion. They reveal imbalances in project collaboration among NUTS3 regions within the study area of this targeted analysis and wider Europe. The eastern parts of the NGZ and some active zones in Europe (e.g. Netherlands and the northern Germany) seem to have a higher number of funded project partnerships, meanwhile Eastern and Southern Europe are surprisingly quiet in terms of such collaborations. The proposed visualization approach could be utilised and developed further for a larger scale trend analysis concerning functionalities of European growth corridors (e.g. TEN-T). In the NGZ, results from such analysis could be used to support networking activities and the development of funding instruments to support interregional project collaborations especially within, and with, the western parts of the corridor, and to legitimize transnational corridor development.

4.2.1 Background and Objectives

This case study conducted analytical experiments with European Regional Development Funding (ERDF) data. The key objectives of these experiments were to analyse interactions between regions in terms of project partnerships with regards to volume and distribution, and to evaluate the usefulness of this dataset for informing policy-making concerning European growth corridors. Another motivation was to produce insights into whether and to what extent the Northern Growth Zone (NGZ) is characterised by project-based networking across regions.

Inter-regional collaboration is seen one of the key factors for regional development and innovation in the EU (e.g. Uyarra et al., 2014). To evaluate such collaborations, a network analysis approach can produce new understandings regarding relations across regions while identifying important actors and potential partners (Merino et al., 2016). This case study analyses such project partnerships, spatially visualising them to better understand the development mechanisms in growth corridor networks. This work builds on previous projects that have conducted similar forms of analysis, for example, the allocation of funding for certain thematic objectives (e.g. Gianelle et al., 2017).

Screening network types and volume can help policy-makers position existing regional projects and their management while prospecting for new ones at various scales of territorial development. Such project network modelling and analysis can facilitate future cooperation and synergies among projects (cf. Harjunen et al., 2017). This case study attempts to validate how spatio-temporal partnerships and network analysis may be utilised by local public sector policy-makers in positioning their funding priorities and, for example, their Smart specialization strategies (RIS3) relevant for growth corridor development (cf. Foray et al., 2012). As one of the aims of growth corridors relates to the connecting role of infrastructure in promoting functional specialisation of areas along the corridors and their division of labour (Kalliomäki, 2012), knowledge on spatial connectivities of projects becomes important to further policy development. Such knowledge on project networks also can be enhanced by other kinds of network data – such as industrial and R&D networks – to broaden the picture of corridor functionalities. For example, a study exploring the connectivity of maritime company board members along the NGZ revealed strategic connectedness along the corridor (Kalliomäki et al., 2018). In addition, an analysis of R&D networks in the corridor revealed a partial disconnectedness of refurbishments from typical R&D networks in the maritime cluster, raising this as a topic requiring further attention in future policy-making supporting network development (ibid.).

A spatial approach to analysing partnership networks thus helps stakeholders position and plan projects related to territorial development. Visualization of project partnership data is a useful tool for identifying groups, assessing their relevance, and analysing partnership evolution. (Merino et al., 2016.) In this respect, this case study also helps reflect the regional project portfolios to other environments and utilise this information to regional specialization (e.g. Gianelle & Kleibrink, 2015, 4; Mariussen et al., 2016, 3). Project management capacities increase through the creation and exchange of knowledge, best practices, pilot actions, etc. (Astrov et al. 2018, 247). By investigating networks more closely, policy-makers and partner-seekers can more effectively assess the specific needs and possibilities for ERDF programs. Altogether, drivers of regional development, prosperity and vitality start from smart strategic positioning and critical choices that lean on regional strengths. These strengths can be developed through inter-regional collaborative relationships. Furthermore, cross-border collaboration and readiness to cooperate are seen as key drivers of regional innovation (Mariussen et al., 2016).

Naturally project partnerships are ever-evolving networks (e.g. Ahn et al., 2011), and more in-depth research is needed to map their temporal dynamics and observe their hidden potentials in a growth corridor context. As evidence-based decision-making relies on measurable variables, there is a need for quantitative evaluation criteria and indicators that would take into account the nature of the projects and their geographical formation more coherently (cf. Mavrič & Bobek, 2015). This was, however, out of the scope in this experimental case study, thus requiring further investigation.

4.2.2 Materials

This case study utilised European Regional Development Funding (ERDF) data downloaded from the keep.eu portal. This database provides public access to details about projects funded by ERDF programs including the partners involved, project budgets, thematic categories, and location. The geographical scope of Keep.eu is wide, aggregating data about all projects and beneficiaries of cross-border, trans-national and interregional cooperation programmes within EU member states and neighbouring countries. The portal also features temporal depth and contains data for three programming periods 2000–2006, 2007–2013 and 2014–2020. The data production and management processes have evolved over time, and today monitoring systems automatically update data to the keep.eu database. These automatic data-deliveries are sent in periodically and the exact timing is decided by each program. Even still, this automated updating made the current programming period much more up-to-date compared to earlier periods. At the time of downloading it had data for 50 percent of all territorial cooperation projects while the previous reporting periods were less comprehensive.[27] This dataset was selected for this analysis. It had data for 2,353 projects with 18,318 partners across the EU.

The key corridor specific policy contexts – land-use, transportation and economic development – identified by this targeted analysis's stakeholders are also emphasized in the literature as highly relevant for corridor development (e.g. Zonneveld & Trip, 2003; Kalliomäki, 2012). Therefore, these themes served as the basis for further filtering the data for this case study. Out of 42 thematic

[27] A detailed search can be found following this link to the keep.eu website: https://www.keep.eu/keep/search/link/Kw4xvXwtVR

categories available in the database, 10 were used and grouped according to the three policy contexts:

1. Land use – Regional planning and development; Rural and peripheral development; Urban development,
2. Transport – Transport and mobility; Logistics and freight transport; Multimodal transport; Improving transport connections), and
3. Economic development -- Clustering and economic cooperation; Labour market and employment; New products and services; SME and entrepreneurship.

During pre-processing, the NUTS3 geographical level was selected as the appropriate level of detail about the network and its functions. The body of filtered material contains 950 NUTS3 areas, 11 000 project partners, resulting 450 000 partnership pairs. Overall, the details that were available on all projects were budget information, thematical division, location of the partners, and type of partner (university, company, government etc.).

Because the data is a depository of reports from the project partners, an extensive processing effort was required to prepare the data for analysis. The selection and vetting process to assemble the dataset and its components was time intensive. These and additional aspects related to the utilisation of this dataset are described in Table 4.3.

Table 4.3. Dataset typology applied to a key dataset used in Case 2.

Dataset European Regional Development Funding (ERDF) data	
Data Characteristic	**Attribute**
Availability	*Open data, through interface*
Level of Processing	*Highly processed*
Intended Audience	*Humans*
Observational Qualities	*Direct observation*
Level-of Detail	*Rolled up*
Level of Structure	*Highly structured, but gaps found*
Refresh Frequency	*Updates weekly, programming data quarterly*
Confidence in Updates	*High confidence* *(Keep.eu is updated constantly, but the data is added to the database a few days after delivery from the programs)*
Extraction Effort	*Requires some effort*
Clarity of Ownership	*Ownership is clear, it is property of the Interact III*

4.2.3 Methodology

The starting point for analysing the dataset was to select and filter relevant ERDF data and retrieve it from the keep.eu portal. This dataset was downloaded in a spreadsheet format and processed so that projects linking spatial areas became more evident and the data was prepared for interactive mapping.

The downloaded spreadsheet had one row per project and NUTS3 spatial area. From this source spreadsheet, a new dataset was generated by the research team based on a comparison of projects and NUTS areas and featuring a pairing of all projects and their respective NUTS3 areas. These processes aimed toward visualising the data as connections between spatial regions on a map.

To further prepare the data for interaction mapping, an extensive manipulation of the dataset was conducted to produce an interaction matrix. The following steps were taken:

- A separate table, pairing project acronyms and NUTS3 areas was created. From this, all instances of several mentions of the same NUTS3 areas were deleted, leaving 10 875 partners. From this all projects with only one NUTS3 area (without partnerships to other NUT3 areas) were removed, leaving us with 10 723 partners.
- The NUTS3-areas were isolated to a separate table which contained 950 unique NUTS3 areas. These areas were combined into pairs, resulting in 450 775 pairs.
- The two tables were combined to see, how many times certain areas were in contact with each other, resulting in a matrix of approximately 18 000 pairs. These pairs were called *collaborationships.*
- Coordinates for the NUTS3 areas were added to prepare the dataset for mapping.

The cooperation and consortium structure aspects were the most convenient to use for pairing of the cells, labelling them, and grouping the beneficiaries. Moreover, as previously mentioned, the ERDF data was filtered to match the focus of this case study with the stakeholders' needs. The approach described above ignores, however, the number of partners within a project for a single area, i.e. a project connection with three partners in one area is treated the same as one with one partner an area. The final part of the analysis method was to visualise the processed data on maps, which was accomplished through an iterative process aimed at producing maps capable of clearly conveying the density of collaborative relationships across regions.

4.2.4 Results

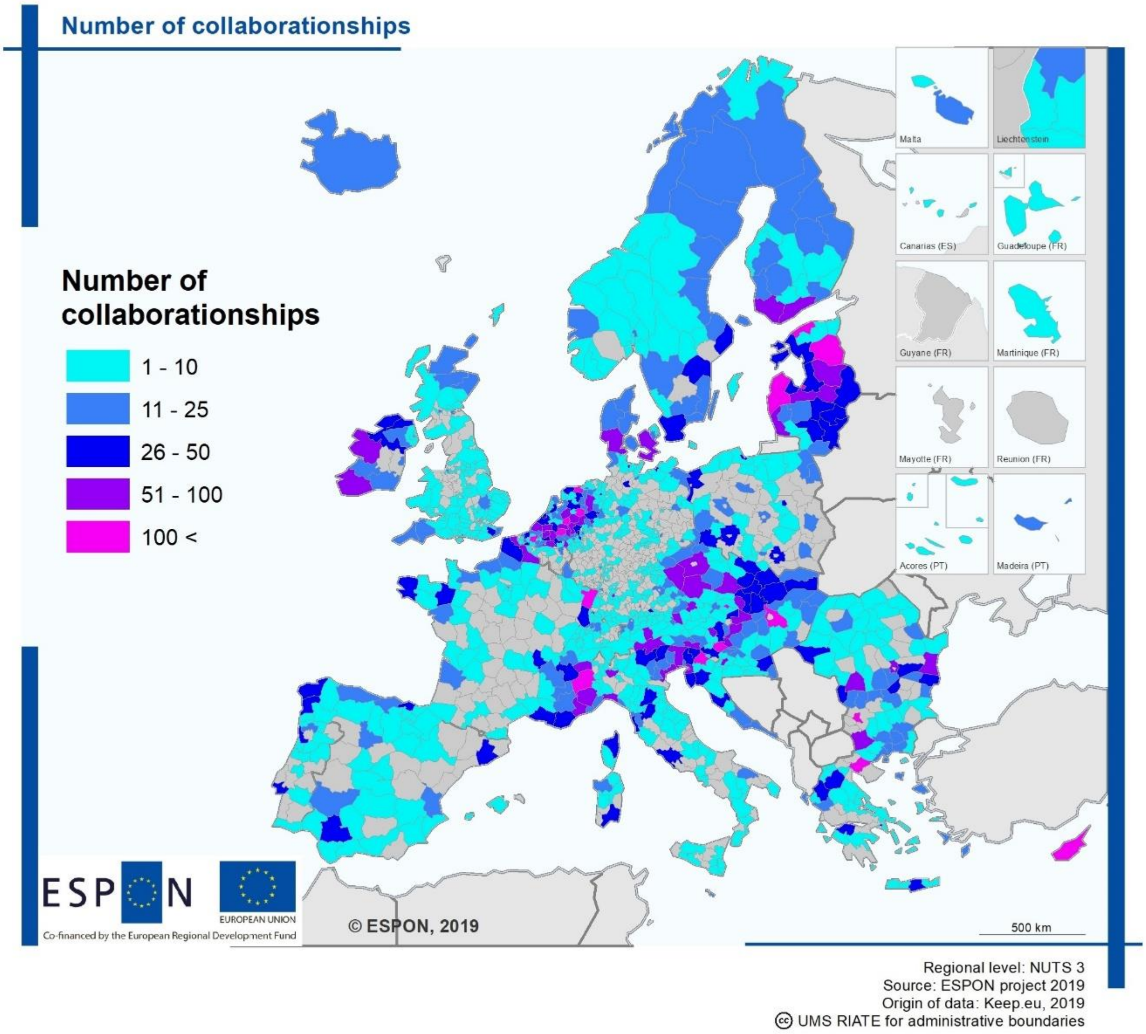
The results of this case study show that the current EU regional development project (ERDF) data provides valuable information about the regional profiles of project activities as well as imbalances in project networks and collaboration activity. The following visualisations present a selected focus based on the research project's stakeholder needs analysis which further links to the proposed growth corridor development framework. The time period is 2014 to 2020. It should be noted that these presented maps do not aim to describe all ERDF project partnerships.

Collaboration activity at the EU scale

Map 4.2 shows the number of project partners per NUTS3 region across Europe. The map depicts the regional activity and strength of the activity across Europe in this case study setting. The analysis reveals a clear imbalance between active and passive NUTS3 regions. Baltic countries are highly active when it comes to the number of collaboration relationships (or *collaborationships*), as well as some isolated NUTS3 regions across Europe. However, a zone from Spain, through France, Germany, and Poland is underrepresented in collaboration activity as well as for example Southern Italy.

Further study is needed to investigate the dynamics and explanatory factors of the activity, as this analysis presents a static picture of the 2014-2020 ERDF programming period and does not take into account for example the urban-rural dynamics and population density across Europe. However, a link has been built in the literature between higher collaboration capabilities and innovation (e.g. Blomqvist & Levy, 2006), which highlights the importance of any vantage point that supports understanding inter-regional collaborative activity.

Map 4.2. Number of collaborationships *per NUTS3.*

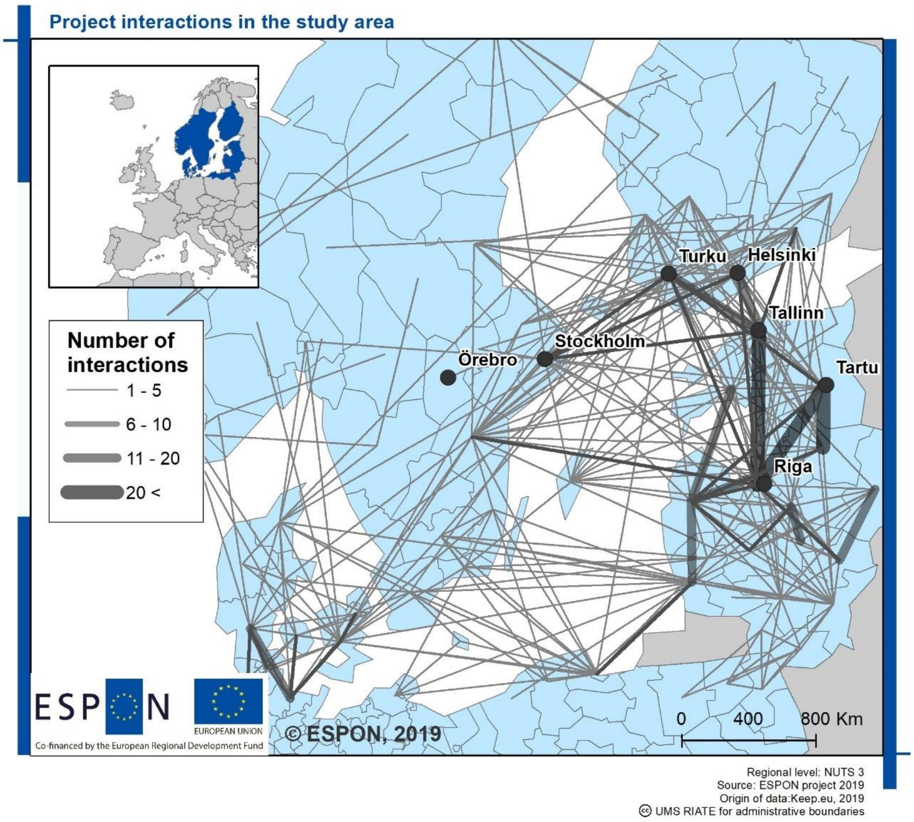


Collaboration in the study area

Next, project interactions were reviewed at the study area focusing on increasing understanding about the collaborative activity and connectivity of the project stakeholders. The NGZ project activity is measured by two metrics. The first one is the project interactions within the study area (see Map 4.3) and the second is the total number of *collaborationships* (see Map 4.4). In addition, the project interactions in selected thematics is presented in Map 4.5 to provide more detailed understanding about the nature of interactions in themes relevant to corridor development.

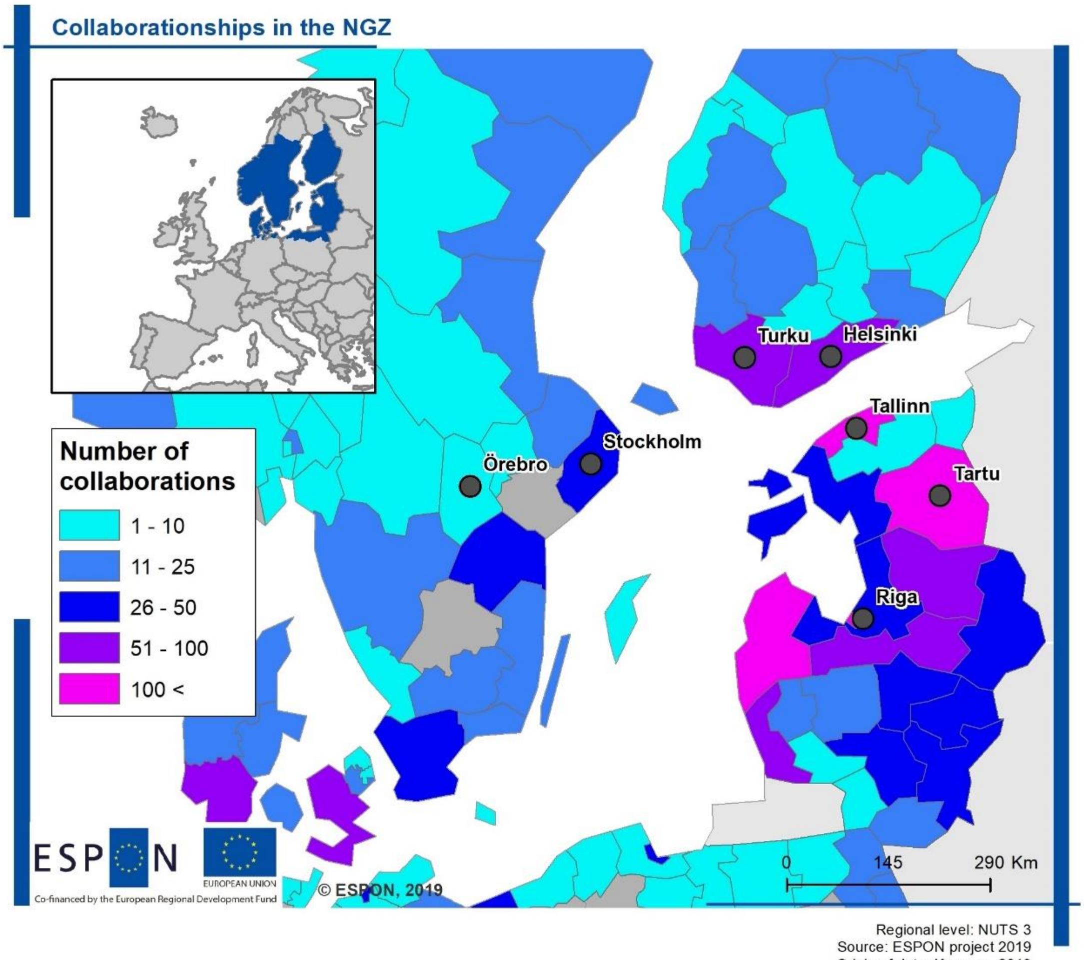
As seen in Map 4.3 the countries across the study area are well linked; the stakeholder countries of this project, Estonia, Finland and Sweden as well as Denmark and Latvia generate the majority of project collaboration. However, the results also show that there seems to be active collaboration in the eastern parts of the study area and the NGZ (high level of *collaborationships* within the Baltic area and between Finland and Estonia), whereas collaboration between Finland and Sweden is not that visible in project interactions. When it comes the international strategic framing of the NGZ, link between Oslo and Örebro as well as even between Örebro and Stockholm is surprisingly passive, which can be only partially explained by the requirements coming from the ERDF. Altogether, it is clearly observable from Figure 4.4. that the project participants tend to be clustered. Comparing the project interactions in depth across the EU's TEN-T corridor network would open up interesting avenues for regional development in the future as a link is often assumed between physical infrastructure and collaboration (see Kalliomäki, 2012).

Map 4.3. Project cooperation links within the study area.



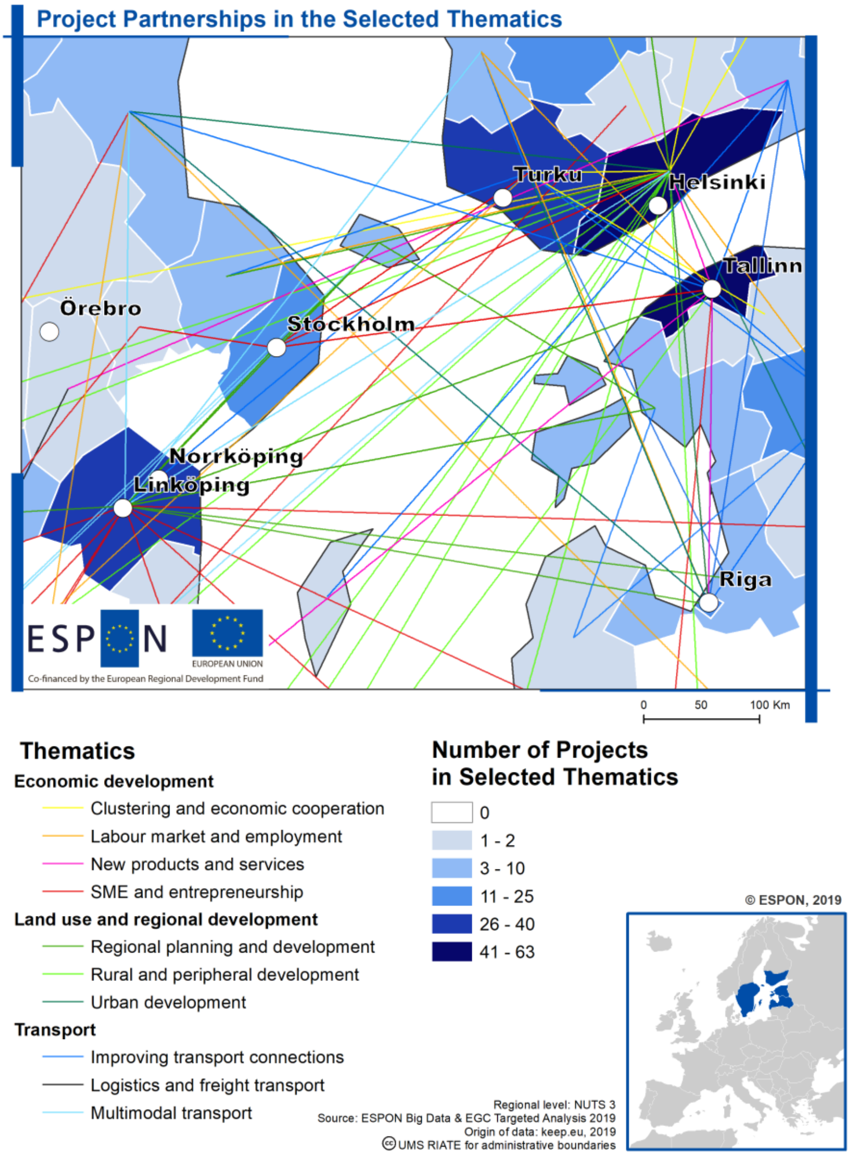
Next, the analysis targeted the connectedness of the NUTS3 regions by selected three themes (transport, land use, economic development) that were seen as most relevant for corridor development. These were observed via the volume of projects. In Map 4.5 the main themes are broken down to their respective sub-thematics. The number of projects and their weight is expressed with a blue gradient in the map. Concerning the study area of the project, the data shows that the project interactions are versatile representing various sub-themes relevant for NGZs development, thus forming a solid basis for corridor-based collaboration.

Map 4.4. Total number of project participants in the NGZ framework.



Regional level: NUTS 3
Source: ESPON project 2019
Origin of data: Keep.eu, 2019
© UMS RIATE for administrative boundaries

Map 4.5. Project Partnerships in selected thematics.



Altogether the role of capital regions seems to be strong in the overall project activity. Tallinn, Helsinki and Turku regions are represented as active partners. Also, the Map 4.5 depicts Norrköping and Stockholm, Helsinki-Uusimaa region and Riga as the most prominent districts by volume in the selected sub-thematics, but again Örebro seems disconnected from corridor-relevant activities. In Map 4.5, Örebro is partnering in only two projects in the selected thematic categories during the 2014-2020 period, which together with other results indicates a weak connectivity to corridor-based development, when reviewed from the perspective of ERDF.

In summary, the main policy-relevant insights produced in this case study were:

- The NGZ and the Baltic Sea region are quantitatively well represented in project activity and partners' connectedness in the EU.

- There seems to be active collaboration in the eastern parts of the study area with a high level of *collaborationships* within the Baltic area and between Finland and Estonia, whereas collaboration between Finland and Sweden is not that visible in terms of project interactions.
- The interlinkages between the TEN-T corridor network and the number of *collaborationships* and the overall project activity should be further studied as data shows that the NGZ/BSR gathers project interaction as well as the central European region. This kind of analysis would further support the development of European growth corridors.
- A lengthy zone from Spain, through France, Germany, and Poland is underrepresented in collaboration activity in the project thematics selected by this study. Also, Southern Italy has few *collaborationships*. This finding should be further investigated e.g. in relation to the urban-rural dynamics in Europe to understand its policy implications.

Based on these insights, further implications for policy can be further discussed.

4.2.5 Implications for policy

The experiments conducted in this case study indicate that big data analysis can reveal potentials for more integrated corridor development and cohesion in the EU when collaborative relationships and capabilities are reviewed in light of strengthening development potential. The results revealed an imbalance in project collaboration among NUTS3 regions, also within the study area of the targeted analysis. Network data visualization is useful in generating an overall view of the project cooperation between NUTS3 regions. The number of project partnerships is also a tool for regional self-reflection. Stakeholders can use these tools to prioritize, monitor and plan regional collaboration more strategically e.g. to support corridor development.

The analysis showed that the NGZ and the Baltic Sea region are quantitatively well represented in project activity and partners' connectedness in the EU. Furthermore, there seems to be active collaboration in the eastern parts of the study area (high level of collaborationships within the Baltic area and between Finland and Estonia), whereas collaboration between Finland and Sweden is not that visible in project interactions. This finding could be utilised to support further policy development when it comes to strengthening the transnational dimension of the NGZ that has so far been a strategic initiative focusing mostly on strengthening the interregional collaboration between the Finnish regions. There is clearly a need to strengthen the transnational collaboration in the corridor to legitimize collaboration under this strategic framework. In this light, the whole idea of a functional corridor can be clearly questioned, thus highlighting the need to utilise complementary data sources related e.g. to transportation of people and goods, which most likely provides a different view on the area's functionality.

The analysis also revealed that southern Europe, especially Italy, France and Spain are notably underrepresented in their number of collaborative relationships along land-use, economical, and transportation domains. Reasons for this should be studied further to understand its policy implications. Altogether passive regions might benefit from activities directed to strengthening collaboration capabilities. The visualization may reveal imbalanced activities or development gaps in strategically important fields for the corridor context. Thus, the proposed visualization approach should be utilised and developed further for a large-scale trend analysis concerning the European growth corridor development. In the NGZ, the results could be used to support networking especially within, and with, the western parts of the corridor.

The outcomes of this case study suggest that the interlinkages among spatial areas along the TEN-T corridor network in terms of the number of *collaborationships* and overall project activity merit further studies. This suggested future research direction is based on the produced insight that the research area of the targeted analysis is characterised by greater project-related interaction along the TEN-T networks. Such analysis would further support the development of European growth corridors and provide evidence to policy-making regarding instruments that foster such collaborations and targeted improvements to existing physical infrastructure.

One viable approach to deepen this approach to network analysis could be to include data about the project budgets in the study area. Future research could, for example, check whether the economic scale of a region and observed regional emphasis have some correlation to the budget size of ERDF projects. If a correlation is found, it would support the arguments suggesting that economic scale and activity levels are linked in respective regions in Europe (Combes, 2003). In addition, there is potential to extend the analysis with machine learning methods and richer semantic representation of various development projects, as well as to combine e.g. diverse funding instruments to the analysis (e.g. through Cordis). In addition, factors explaining the geography of collaboration should be further studied.

Altogether, network analysis of ERDF project data helps in the evaluation of interaction between partners and modelling inter-linkages (Bramoullé et al., 2016). To get a better picture about the project partnership network new sources of data should be harnessed. This would mean that traditional quantitative factors could be enhanced with additional datasets, such as automated transcripts and summaries of development discourses or social network data. These kinds of hybrid analysis experiments to assess project collaboration could provide fresh angles to the extant network analysis literature (cf. Huggins & Thompson, 2017) and territorial policy-making.

4.3 CASE 3: MOBILE POSITIONING DATA FOR AN ESTONIAN EVERYDAY MOBILITY DATABASE

Box 4.3: Summary of Case 3

Objective: To develop a methodology for producing an everyday mobility database which contains OD-matrices of movements between territorial communities.

Data used: OD-matrices are based on mobile positioning data and applied to road network based on Dijkstra routing algorithm and Open Street Map roads data. Mobile positioning data contains locations of call activities (Call Detail Records (CDR)) in network cells (location, time and random unique ID).

Novelty of the approach: The mobile positioning data has high accuracy in time, the data allows to short-term differences (find the OD-matrices by months), include in addition to movements between the place of residence and the workplace, also other regular places and differentiate the movements of different social groups (gender, age, nationality) and the types of movers.

Challenges: The main limitation of passive mobile positioning data is access to data, because mobile network operators are hesitant to provide their data and relatively long value chain of implementing mobile positioning data, which requires expertise from several research fields.

Policy implications: The resulting database would support mobility-related policy-making.

4.3.1 Background and objectives

For this case study, Mobility Lab of University of Tartu partnered with the Estonia Ministry of Economic Affairs and Communications of the Republic of Estonia. The objective of the Ministry is to develop a high-quality database of mobility and traffic data covering all of Estonia. This database would be an important data input for the ministry in spatial planning decision-making processes and serve as a way to answer questions related to transport and mobility. The aim of this case study is to develop and implement a methodology that can establish and maintain an everyday mobility database for Estonia based on mobile positioning data. A requirement for this database is that it is compatible with Estonia's existing mobility databases.

The source material for this project is mobile positioning data provided by a mobile operator. Mobile positioning data provides more accurate and higher resolution spatial and temporal information about individual-level mobility patterns than conventional methods. Mobile operators collect this data continuously and it has a fine-grain temporal resolution. Passive mobile positioning has two main strengths – historical depth and extensive sample. Mobile network operators continuously collect Call Detail Records (CDR) and these datasets can cover several years. For example, the Mobility Lab at University of Tartu has already gathered a continuous 12-year time series of CDR data from Estonia. Globally, more than 5 billion people in the world have a mobile phone and the number mobile subscribers is increasing (GSMA Intelligence 2018).[28] This means the method developed in this case study to produce the everyday life database could be applied anywhere. For this case study, raw data acquired from Telia were processed to produce OD-matrices of movements between territorial communities based on mobile positioning data.

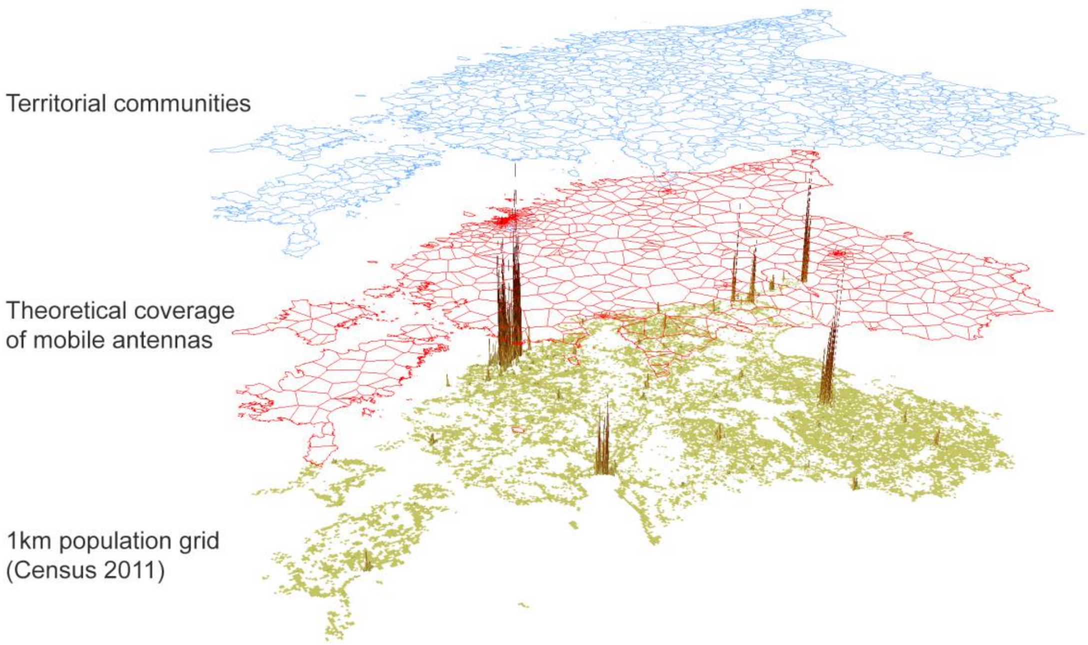
The usage of mobile positioning data in mobility studies has been increasing over the years. Mobile positioning data is well-suited to provide spatio-temporal insights concerning locations, flows, and spaces. This data can be analysed at individual or aggregate levels. Mobile positioning data reveals locations that an individual visits and important activity locations such as anchor points – e.g. home or work) (Ahas et al., 2010b). This enables an examination of both individual spatio-temporal mobility between activity locations (Ahas et al., 2007b) and an individual's daily and long-term activity spaces (Järv et al., 2014).

[28] For live updated estimate of mobile phone subscriptions, see GSMA Intelligence (2018), https://www.gsmaintelligence.com/ (Accessed 8 February 2019).

The scaling up of individuals' activity locations and movements reveals dynamics of human presence and mobility for whole populations within spatial structures. Aggregated individual mobility data can, for example, uncover daily work-related commuting flows (Ahas et al., 2010a). Aggregated activity locations and flows unveils the urban spatio-temporal structures, land use and settlement hierarchies of societies and can shed light on how these structures change in time (Ahas et al., 2015; Louail et al., 2015; Pei et al., 2014; Silm and Ahas, 2010). This helps to identify and monitor hinterlands of cities, functional urban regions, and growth corridors (Novak et al., 2013). Mobile positioning data have been widely used in transportation and mobility studies (e.g. Calabrese et al., 2010; Isaacman et al., 2011; Yuan et al., 2012) as well as in geographical research (e.g. Blumenstock and Fratamico, 2013; Silm and Ahas, 2014a). In Estonia, mobile positioning data has been used to map the commuting areas of municipalities by order of the Ministry of the Interior (Ahas et. al, 2010; Ahas & Silm, 2013).

For this case study, Mobility Lab prepared a dataset for each month from January 2017 to April 2018. Based on their earlier experience and preliminary analysis of the data, the team chose to target a spatial granularity at a Territorial Communities Layer (Figure 4.3). The size of each area in this layer varies from urban to rural. As the mobile network is very dense in Tallinn the spatial accuracy can probably be higher. The completed datasets will be in machine readable format (*.csv) and available directly from Mobility Lab website (http://aasa.ut.ee/OD).

Figure 4.3: Geolayers used for modelling of the Origin-Destination matrix.



4.3.2 Data & methods

The main data source for the everyday mobility database developed in this case study is passive mobile positioning. Passive mobile positioning data is automatically logged and stored by mobile operators for call activities or movements of handsets in the network. For this case study we used a dataset containing locations of call activities, or Call Detail Records (CDR) in network cells. This data includes locations, times and randomly generated unique IDs for every handset in the network. The spatial resolution of this dataset varies by the size of network cells. Other data characteristics are described in Table 4.4.

Table 4.4: Dataset typology applied to a key dataset used in Case 3.

Dataset Call Detail Records, Telia (in Estonia)	
Data Characteristics	
Availability	Purchasable Proprietary Data
Level of Processing	Raw
Comprehensiveness	Partial, can be increased by adding CDR data from other carriers.
Intended Audience	Machine Readable
Observational Qualities	Direct Observation
Level-of Detail	Fine-Grained
Level of Structure	Highly Structured
Refresh Frequency	Monthly
Confidence in Updates	High Confidence (updates are highly likely to occur)
Extraction Effort	Requires some effort to prepare data for use
Clarity of Ownership	Ownership is clear and singular
Spatial Resolution	Local, sizes of network cells vary
Temporal Resolution	A range from minutes to years

Privacy concerns

To address privacy concerns, the raw data is pre-processed and the records are made pseudonymous so they do not contain any back-traceable personal information about the user of the phone. Before providing data to the research team, the mobile operators aggregate the geographical data in their log files and take steps to anonymize the data. For these reasons, researchers can use the data they acquire from the company for surveys for scientific purposes.

To recognize individual-level mobility, which is essential in order to analyse repeat visits and routine mobility, a randomly generated unique ID number is assigned to every phone. Every respondent is given a unique ID, a numerical pseudonym that remains constant for every contract in the system, thus making it possible to analyse the digital trail of a phone's physical presence in time and space.

The collecting, storing and processing of the data used in this case study complied with European Union requirements regarding the protection of personal data according to EU directives on handling personal data and the protection of privacy in the electronic communications sector. Separate approval for the development of this database was also sought from the Estonian Data Protection Inspectorate.

Data

The data are collected and processed by Telia, Estonia's largest mobile operator with approximately 45% of the market share, the Mobility Lab in the Department of Geography at the University of Tartu and a spin-off company named Positium LBS. The structure of the dataset is shown in Figure 4.4.

Figure 4.4: Structure of the mobile positioning dataset.

data table

parameter	value
ID	246513389
event	call
timestamp	12:15:11 07/04/2014
cell ID	6547

antennas table

parameter	value
cell ID	6547
longitude	25.80527
latitude	58.34998

Within this study, the meaningful places for a respondent originate from the use of the anchor point model, which was developed by the Mobility Lab at the University of Tartu and Positium LBS. The model consists of eight steps that include determining the cells visited, cleaning the data and determining the anchor points (Ahas et al. 2010). An anchor point is defined using the concept of actual activity locations at which people regularly stay or visit and from where they make calls or sending texts. One way the model does this is by converting the locations of outbound call activities into meaningful places (Ahas et al. 2010). Anchor points are determined based on the location and the timing of the call activities of each user over a one-month period (Ahas et al. 2010). Therefore, the model helps to assign locations that are meaningful to mobile phone users for every calendar month. On average, approximately 420 000 active respondents per month were identified whose home anchor points could be defined using the anchor point model (with a varying maximum of approximately ±10%).

This study uses all types of anchors calculated by the model. 1) Home anchor point – an everyday anchor point that is the probable location of the device owner's home. 2) Work-time anchor point – an everyday anchor point where the device owner spends the most time during a workday. The anchor is called a work-time location because it is not possible to differentiate between work, school, and other activities in the place where a person regularly and most often spends time during business hours of a month. 3) Multifunctional anchor point – an everyday anchor point in which the home and work-time anchor points are located in the same network cell and cannot be separately identified. 4) Secondary anchor point – anchor points that have lower visiting regularity than everyday anchor points (Ahas et al. 2010). These anchor points allow us to investigate meaningful locations and people's daily activity spaces as well as more permanent moves such as changes of residence.

The accuracy of spatial location information depends on the structure of the mobile network. The network cell level is the minimum level of analysis. The structure of the mobile network and therefore the theoretical size of each base station are not fixed because the network changes with time due to the setup of new base stations. In Estonia, more densely populated areas such as towns and the vicinities of the main roads are covered by more base stations. In towns (such as Tallinn, Tartu and Pärnu), the spatial accuracy is 100–1000 m, whereas in the remainder of Estonia, the accuracy falls to between 1.5 km and 20 km. The average size of the theoretical service area in 2011 was 43 km^2, and the median was 18 km^2. Phones normally switch to the closest base station or the one with the strongest radio coverage or the best 'visibility' levels. Because the network structure and placement of base stations changes with time, it is important to consider such modifications if the aim is to expand the time span under investigation to several years.

For this case study, all of the locations the respondent visits in at least five days during a month will be added to the OD-matrix. For every user ID one home and one work location is defined per month.

Methodology

Data inputs for current study is coming from mobile positioning data of one Estonian mobile operator. Dataset itself when received from the mobile operator is anonymous. Any attempts to identify the individuals is illegal and – without crossing the principle of reasonable effort – also impossible.

In general, the data is log file of calling activities (taxable calls and SMS'es). This simple dataset contains three variables:

- pseudonymous user ID;
- starting time of calling activity;
- location of mobile antenna which was used for current communication event.

The research team's experience working with this form of data supports the claim that most of people's life is very routine: the majority of time we spend at home, another important location is related with work, and visiting of other places is episodic but does not rule out some forms regularity.

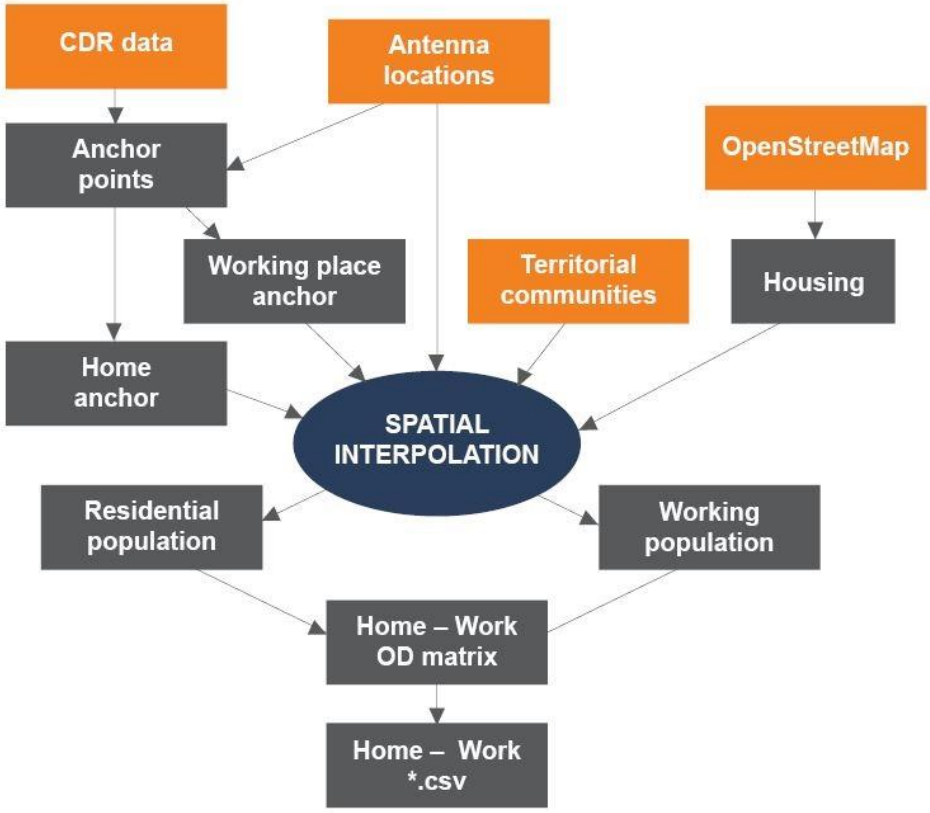
Everyday mobility is described as movement between regularly visited places (home, work, and free time). A methodology developed by Mobility Lab of University of Tartu can detect regularly visited meaningful locations, or anchor points, of phone users. This methodology can detect and distinguish between three different types of anchor points:

1. Home
2. Working time
3. Secondary (Other)

For current study, home and work anchor points are most relevant. Anchor points are calculated for the clients of one mobile operator and on the basis of antenna locations. This means the anchor points have to be interpolated and modelled to the level of meaningful spatial units that can be converted to the whole population. The workflow used to produce the OD-matrices is visualized on Figure 4.5.

Because the source dataset does not depict the actual routes taken by users, all the movements are constructed as routes that are most likely between the users anchor points. As a simplification, all movements start from home, meaning that real travel chains stay hidden.

Figure 4.5: The workflow from mobile positioning data to origin-destination matrices.



4.3.3 Results

Mobile positioning data can be a highly useful data source to study spatial distribution and mobility of a society. In case of mobile positioning data, the spatial accuracy varies from 50–100 m in urban areas to 10–30 km in remote rural areas. This means that also the OD-matrices can't be modelled in higher accuracy. In the case of Estonia, the suitable level of the spatial administrative hierarchy is a territorial community (in Estonian "kant").

Based on previously introduced methodology we calculated monthly OD-matrices of regular movements between all territorial community. This resulted in csv-files which are uploaded to webpage of the Mobility Lab of University of Tartu (http://aasa.ut.ee/OD). The access to the methodology and data is freely available. The datasets are semicolon-separated csv-files that can be opened in any spreadsheet software or imported by databases. These files contain 11 variables:

- Name of the starting territorial community,
- Name of the destination territorial community,
- ID of the starting territorial community,
- ID of the destination territorial community,

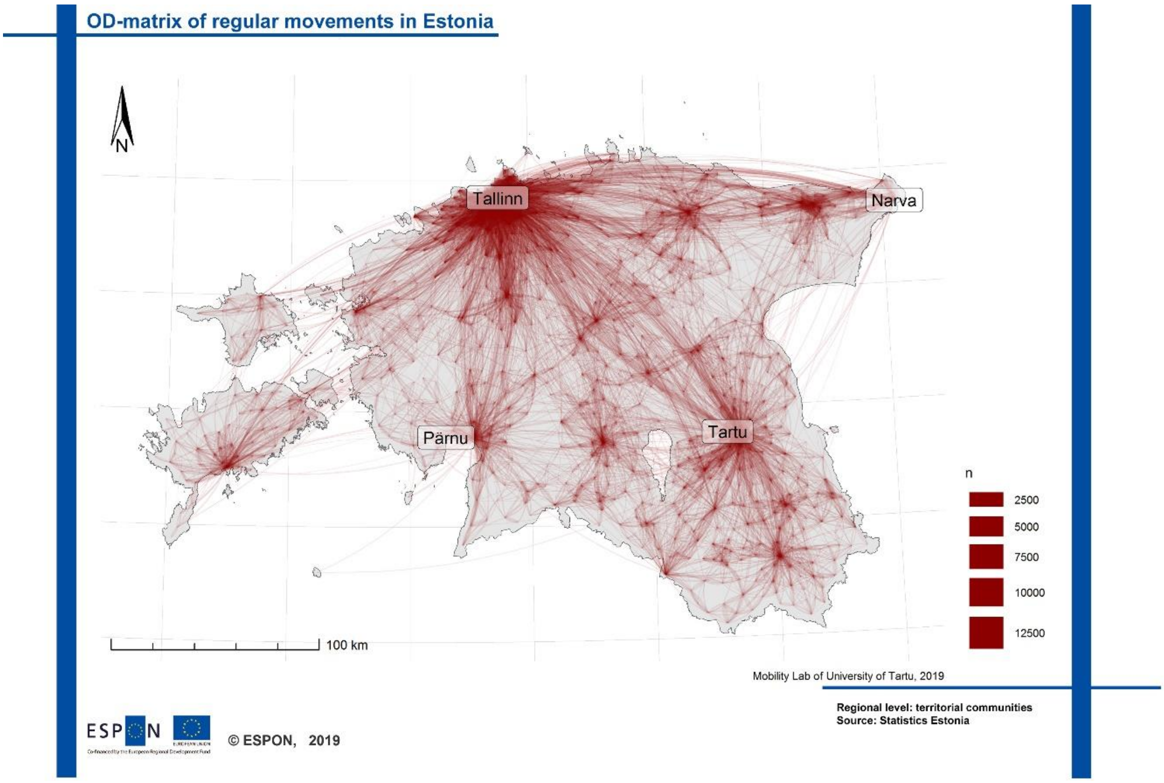
- Geographical coordinates of centroid of territorial community
- Route ID
- Population size in territorial community
- Number of regularly moving people.

The cases where origin and destination of regular movements is the same territorial community reflects the number of people who don't move regularly out of a certain territorial community. Centroid coordinates are projected to Estonian official coordinate reference system (L-EST97; EPSG=3301). Route ID is the unique identifier for every possible modelled route between territorial communities.

On Map 4.4 the whole OD-matrix is visualized. The number of regularly moving people between territorial communities is distinguished by line size and transparency. This dataset supports analysis and description of the spatial relationships in the administrative settlement system. Based on regular flows we can see the settlement hierarchy; distinguishing central places and their catchment areas. Although Estonia is very small, we can even identify the main movement corridors between bigger cities. The fact that an OD-matrix is calculated for every month in the study period makes it possible to analyse them for seasonal effects and differences. If the time series becomes long enough, then changes and trends over time can be quantified. In general terms, this data set can be further analysed to help detect main commuting corridors and estimate the differences in daytime and night-time populations.

This data can help to describe and understand the scale and extent of gravity of central places. Example of the influence of city of Tartu is visualized on map 4.5 and Figure 4.5. The effect where the interaction between locales declines as the distance between them increases is called distance decay. On Figure 4.6 it is also visible, that relatively more people are visiting regularly Tartu and not vice versa.

Map 4.6: Regular movements in Estonia according to the mobile positioning data.



On maps 4.6 and 4.7 the regular movements of people are modelled and summarised for each road segment, and then visualized. Although we don't know the real movement paths of respondents we may assume, that usually they choose the fastest and shortest route. Based on that assumption the OD-matrix is joined with OSM roads layer and movement path from origin to destination is routed with Dijkstra's algorithm. In several cases the preferred route in real life differs from the modelled route. After calibration with real data the dataset could be a good cost-effective alternative for a traffic census and may help to plan and/or monitor public transportation and predict road construction needs.

Map 4.7: Share of population regularly visiting Tartu city.

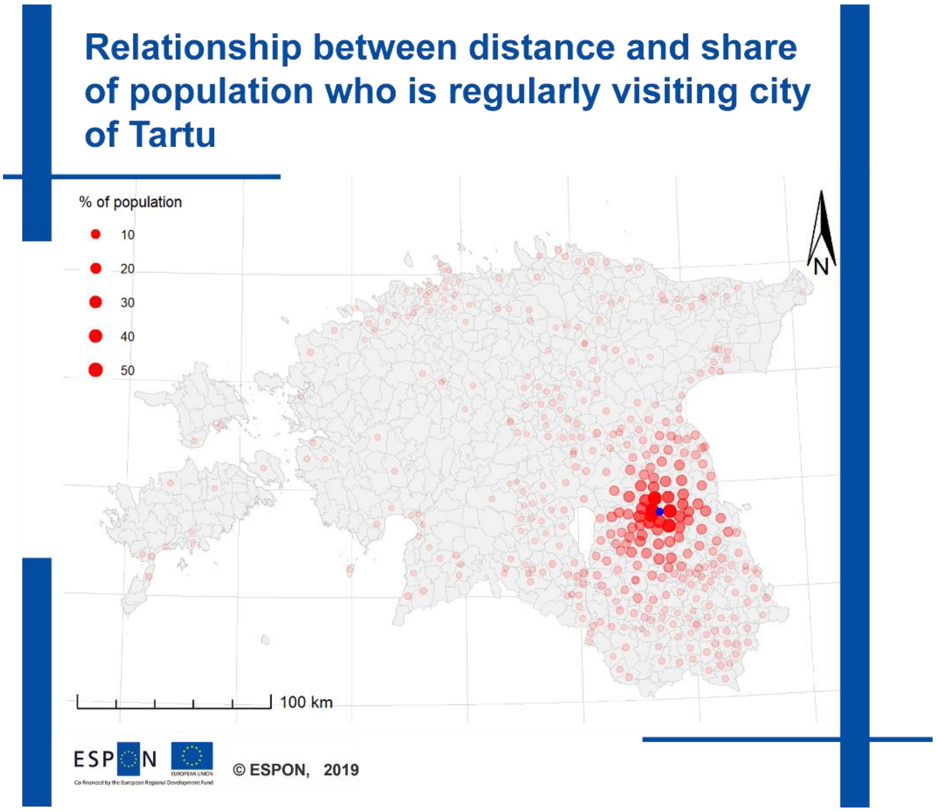
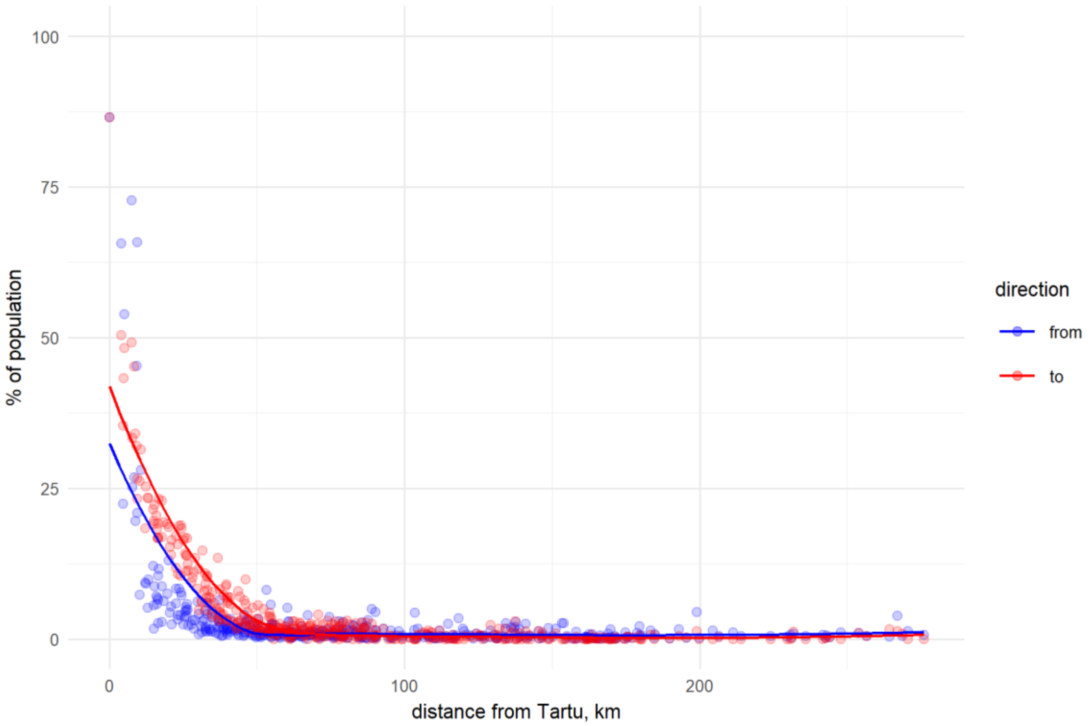
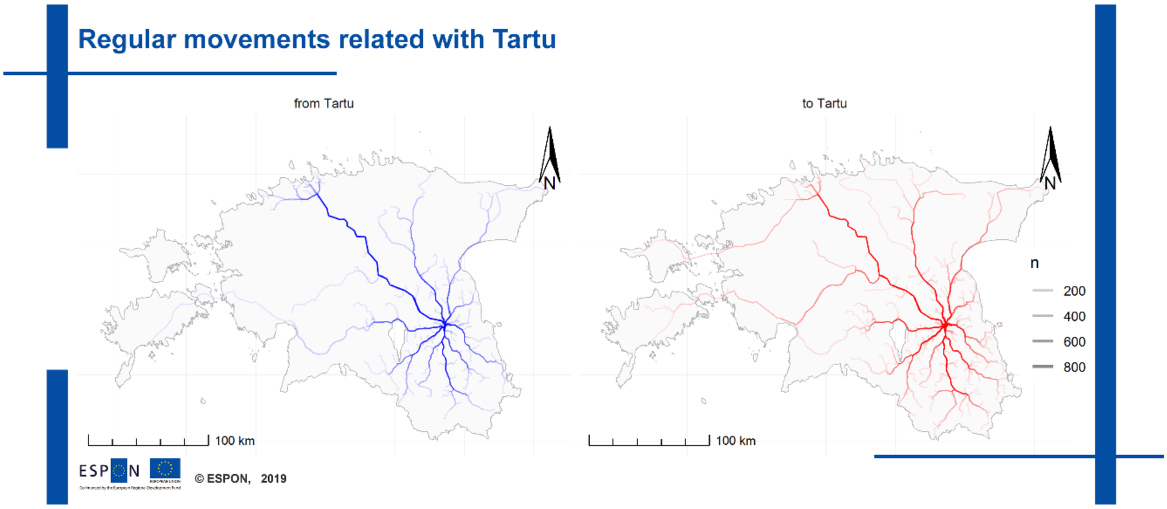


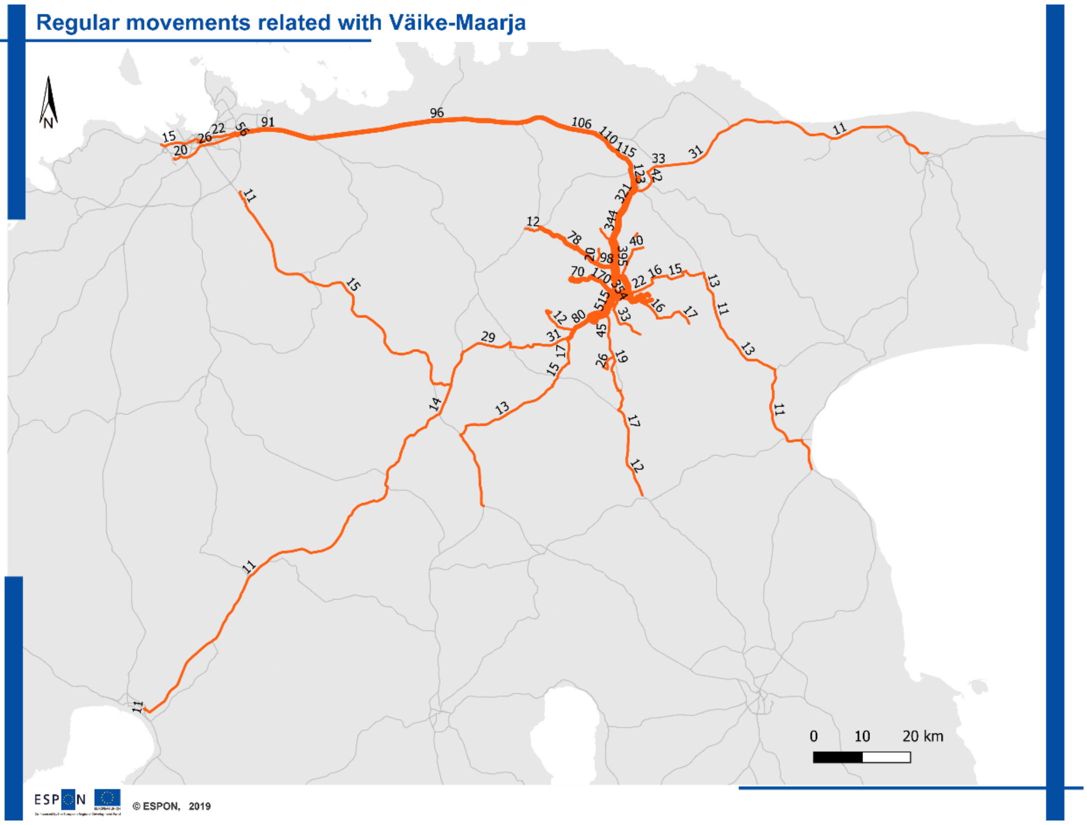
Figure 4.6: Relationship between distance and share of population who is regularly visiting city of Tartu.



Map 4.8: Regular movements to and from with Tartu routed on road network.



Map 4.9: Number of regularly moving people per road segment related with one territorial community (destination: Väike-Maarja)



4.3.4 Novelty

The planned database is innovative in several aspects: level of temporal detail, accuracy, inclusion of secondary anchor points, and modelling mobility to the road networks.

This kind of data was previously only available from censuses and even then, only describing regular movements between home and working place. All the secondary movements have remained out of focus. Additionally, we can argue that life's pace has remarkably accelerated over the years and the 11-year periodicity of censuses is not enough to give a complete picture of the mobility of people.

For example, the start and end of some phenomena (e.g. an economic crisis, urban sprawl, etc.) might occur between two such censuses. It means that the understanding of those processes stays at a very general level. Mobile positioning data allows to estimate the location of the population and the movements more accurately in time (by years and months). Furthermore, other official data sources may also give an insufficient or incorrect view on mobility. For example, up to 25 percent of home addresses registered in the Estonian population registry do not correspond to the reality; while traffic surveys have poor spatial coverage and do not usually give any information about the origin and destination of individual trips.

Another innovative aspect of mobility data is the usage of secondary anchor points which support a more comprehensive view of individual mobility compared with other databases. And thirdly, the OD-matrix will not be just traditional ‘starting point – end point’ table. Instead, individual movements are modelled on real road network.

4.3.5 Additional Possible Applications

In this case study, OD matrices have only been found for movements within Estonia, but mobile positioning data also allows to estimate flows for cross-border movements based on mobile network operators roaming data. Each mobile network operator can provide data related to its own country (domestic, inbound and outbound), so can find movements and directions related to the country, for example, where people go from Estonia and from where people come to Estonia. If similar data would be available for all countries, the same methodology could be used for finding cross-border movements and corridors.

In addition, if there are also social characteristics of the people (gender, age, nationality etc.) in the mobile positioning data, the same OD-matrices and movements corridors can be found for different social groups. OD matrices can also be found for different types of movers (for example, tourists, regular movers, students, etc.). This would give to the corridors also a qualitative dimension. However, this data is self-reported by phone users and may at times be unreliable.

The more carrier data included in a database like this, the more comprehensive the dataset becomes. In the case of Estonia, there are three major mobile operators. So far, this daily mobility database includes data from only one of them. It is possible to imagine a comprehensive Europe-wide everyday mobility database it included Mobile Positioning Data from all mobile operators. There would likely be significant policy and practical hurdles to producing such a database.

4.3.6 Limitations

The main limitation of passive mobile positioning data is access to data, because mobile network operators are hesitant to provide their data. There is relatively long value chains to implementing mobile positioning data, which requires expertise from several research fields at the same time (Ahas et al., 2008a). Besides ethical issues and privacy concerns, lack of ground truth data to validate obtained mobility findings have hindered the implementation of the method so far.

4.3.7 Implications for policy

Mobile Positioning Data is widely recognized as a promising data source for analysing human mobility. In countries like Estonia and Indonesia, mobile positioning data is already used as an input for official statistics. This type of data and the method demonstrated in this case study can reveal mobility patterns in transportation corridors and other mobility infrastructure with high temporal resolution, allowing policymakers to see short-term differences in usage, including seasonal ones. In addition to indicating movements between places of residence and workplaces, other movements to additional locations can also be revealed. To a limited extent, this type of data can provide insights regarding the mobility patterns of different social groups based on device owners' self-reported demographics such as gender, age, and nationality.

The biggest limitations to a wider utilisation of this form of big data in supporting policy-making is due to public fears of losing privacy and obsolete legislation. The wider public does not have a clear understanding of the personal risks and collective rewards linked to sharing this kind of data. Legislation, meanwhile, either does not address the topic of how mobile positioning data should or should not be used or describes these parameters very poorly. To address these issues, nations and the EU ought to clarify – through legislation and public awareness campaigns – under which conditions and in what ways mobile positioning data can be used and benefits society and what assurances are in place to protect individual privacy. Elaborating on this need for greater clarity, this case study informs the following three recommendations to support uptake of this type of data into policy-making:

1. Legislation should provide clear and unambiguous exemptions for mobile service providers or communication undertakers to allow transferring mobile positioning data to government agencies or universities for research and statistical purposes.
2. National supervisory authorities responsible for monitoring the application of GDPR should also provide services to verify whether research is in compliance with GDPR. For example, in Estonia, the pre-GDPR legislation provided the supervisory authority a possibility to verify compliance of a research design. That option was removed for most cases in post-GDPR legislation. Therefore, mobile service providers and third-party researchers now have little to any assurance from authorities that research projects using this data comply with the law, exposing them to legal risks and possible misuse of data.
3. In order to reduce risk to all parties involved in research involving passive mobile positioning data, a specific code of conduct should be developed and approved pursuant to Article 40 of the GDPR, 'Code of Conduct', could provide further guidelines to assess the choice and extent of measures and procedures needed for such transfers of passive mobility data. These guidelines, could for example, indicate to what level a service provider should anonymize data prior to data transfer.

4.4 CASE 4: BIG DATA HACK

4.4.1 Objectives and structure of the Big Data Hack

Hackathons are typically organized as intensive workshops in which teams compete against each other to produce the most innovative and viable solution to a given challenge. Typically, a hackathon aims toward digital innovation; however, recently an increasing number of hackathons are organised to address wider societal issues and social innovation alongside traditional tech-centric approaches (see, e.g., Briscoe & Mulligan, 2014). Hackathons also tend to be oriented toward addressing an existing or anticipated customer or stakeholder need. A key motivation for organizing a hackathon is to explore new ways of doing things using state of the art concepts and technologies. In a sense, hackathons are ad hoc testing environments which require little time and investment from stakeholders, but which can directly inform the development of new initiatives. For participants, a hackathon is an opportunity to solve real business cases and challenges, create unique solutions, network with other participants and potential employers, and enhance personal collaboration skills in multidisciplinary teams.

In the Big Data & EGC project, the Big Data Hack was organized in 17–18 January 2019 at Åhuset, Turku. The hackathon was organized in collaboration with the Open DaaS (Open Data as a Service) project, which aims at creating new business possibilities out of open data. The participants were recruited by advertising the hackathon on mailing lists as well as on the intranet of University of Turku and Turku University of Applied Sciences. In addition, participants were recruited by the Mobility Lab of University of Tartu. The advertisement was written as an open call for proposals for anyone interested in big data and working in multidisciplinary teams. The objective of the call was to attract a variety of expertise from multiple academic disciplines. Students and researchers were encouraged to apply.

A total of 22 students and researchers applied to the event from Finland (n=14) and Estonia (n=8) and all were accepted. The disciplines of the participants ranged from bioinformatics, information system sciences and future studies to economic sociology, geography, visual arts, and cultural studies. In addition, practitioners working in the field of rescue services were involved in Estonia as there was a practical interest to explore the potentials of big data in optimizing prevention work.

To familiarize the students with the concepts and problem space they would be working in, a briefing was held for all registered participants the week before the event. During the briefing session the participants were told about objectives of the ESPON 2020 programme and how the hackathon related to the Big Data & EGC project as a whole. The participants were also given a pre-assignment of identifying potential data sources that could be utilised to describe urban connectivity from a new perspective. The pre-assignments were then used as a basis for organizing the hackathon teams.

The purpose of the hackathon was to propose approaches for generating policy relevant insights regarding urban connectivity via new data sources and new combinations of data sources. The teams were encouraged to investigate potentials of big data in promoting interregional and international collaboration across cities and diverse social and economic agents. To support this objective, links to several open datasets were provided to the participants as well as special access to private datasets. Open datasets included automated traffic measurement dataset from Case 1 (see chapter 4.1) and project network data from Case 2 (see chapter 4.2) while privately owned datasets including Futusome Intra and Suomi 24 which contain social media posts and discussions in Finnish. In addition, the participants had were encouraged to use datasets they had found on their own or explore potentials in other open data, in line with the main objective of the Open DaaS project.

The participants were divided into three teams, two of which worked from Turku and one of which worked remotely from Tartu. The teams self-organized themselves with the help of the research team. One team in Turku focused on collecting and combining big datasets to produce an analytical tool while the other Turku team worked mostly on a conceptual level considering how combinations of existing datasets and ones that could be made available if so chosen can be combined to produce more comprehensive insights into social commuting. The team in Tartu focused on finding new data sources to optimize fire prevention work in Estonia. At the end of two days of intense effort, the teams presented their work to the research team and stakeholder representatives, answered stakeholder questions, and received some initial feedback. The teams were then given approximately two weeks to refine their ideas and create pitches for them.

4.4.2 Team Results

The teams pitched their ideas to experts and stakeholders on 31 January 2019 (see Box 4.2). This panel consisted of representatives from the Regional Council of Southwest Finland, City of Turku, and local companies from Turku region. In contrast to most hackathons, Big Data Hack was not organized as a competition. This decision was made because the hack was about exploring unknown data-driven possibilities instead of providing viable solutions for a well-defined customer problem. Thus, the outcomes were also evaluated according to their viability and potential effectiveness in promoting the ESPON Big Data & EGC project’s objectives.

Box 4.4: Outcomes from three teams of the Big Data Hack event

Team A: Using big data for social commuting. This team had expertise in information system sciences, maps and geography, visual arts, behavioural science, and future scenario planning. Their work was based on creating added value for regional planners by combining different open datasets (social network data, social media data, and different location-based data) with potential to generate insights regarding commuting patterns and supporting the development of a sharing economy. They proposed such an approach would generate high-resolution analyses covering physical, social, and digital aspects of commuting to support new social commute solutions.

Team B: Data-driven ecosystem service for population flow optimization. This team proposed that by combining existing datasets describing, e.g., transport, demographic and economic data, the flow of people can be analysed, and forecasts of individual mobility can be made. Deeper understanding about intercity flows would not only help in evaluating past infrastructure decisions but also support understanding current behaviour while providing tools for future decision-making. The ecosystem would combine datasets and activity from public and private actors to work towards more sustainable solutions. The improved understanding of flows could be utilized e.g. in land-use and transportation planning as well as more efficient resource allocation and service provision.

Team C: Big data for preventing fire deaths and optimizing prevention work in Estonia. This team aimed to develop a data-driven approach for identifying additional fire-risk factors that could be used to locate high fire hazard areas in order to optimize fire prevention work. In the current situation, the are some inaccuracies in the data because of the issues related to classification, lack of capacity to analyse data, as well as restrictions regarding access to the data. The team proposed that combining additional datasets, such as crime data and ambulance data, with existing ones, would provide new insights upon which to make fire prevention interventions.

From the perspective of European growth corridors, the results of the hackathon explored several ways big data could be used to provide insights into corridor functionalities that could be used by policymakers to guide corridor developments in socially beneficial ways.

For example, many regions and cities have goals to make their mobility systems more sustainable. One pathway toward that goal is to promote carpooling. When it comes to questions related to social commuting and increasing carpooling for public value, corridors can become high-impact platforms for piloting social commuting solutions. In the idea developed by Team A, social media platforms and related data could would be leveraged to encourage people living near cities to carpool. Various data sources related e.g. to physical locations of commuters and their mobility patterns as well as mobility preferences and habits (see, e.g. Malokin et al., 2015) could be combined with existing commuting data focusing on individual commuters and their places of residence to broaden. The resulting insights would then broaden the range of interventions possible to promote carpooling. For example, several studies analysing carpooling behaviours have emphasised the role employers can play in promoting sustainable commuting patterns by supporting social commuting (e.g. Vanoutrive et al., 2012; Neoh et al., 2017). By combining rich social data with other sources, additional leverage points could be identified. Using such insights, collaborative corridor governance could provide a platform for public-private collaboration in the field.

Using big data to inform policy-making in growth corridors will require new tools for gaining new insights. Team B attempted to fill this gap by proposing a data-driven ecosystem for flow optimisation in growth corridors. The need for developing a data-driven ecosystem and collaborative data governance has been a key driver of sustainability initiatives at various scales of spatial development (e.g. Chen & Lee, 2018) as for example MaaS practices rely on increasing data integration (Milne & Watling, 2018: 7, citing Ambrosino et al., 2016 and Heikkilä, 2014; see Box 3.2 above). In addition, the utilisation of mobile positioning data and tools such as Telia's new Crowd Insights service that provides different tools related to trips, flows and locations also feeds into flow optimization when combined with other data sources.

Team C's idea related to optimizing fire prevention work highlights how improvements are needed not only in how physical infrastructure is maintained but also in how efficiently data is utilised across administrative boundaries. Their idea relates, for example, to a study made using predictive data analytics in Pittsburgh to predict fire risks in the metropolitan area which addressed challenges of insufficient data sharing across municipal agencies (Levine, 2018). It also relates to a study by Wamba et al. (2015) which highlights the importance of a robust platform to handle multiple data sources for improved emergency service management. As multiple agencies have different kind of structured and unstructured data, the combination of these current datasets with historical statistical data could improve service delivery and preventative interventions. Because growth corridors provide the physical backbone for the prevention work across administrative boundaries especially in cross-border areas, optimizing prevention work should be also addressed at the corridor development scale instead of limiting analysis to jurisdictional boundaries. As an example of a concrete resource, a dashboard to display real-time data (e.g. weather conditions) has been used in the case of Australian state emergency service to coordinate the responses in emergency events, which is then routed via various channels (e.g. radio, phone, social media) to dedicated stakeholders (Wamba et al., 2015).

4.4.3 Lessons learned from the hackathon process

The hackathon proved to be an interesting method for exploring potentials of big and open data in exploring corridor flows, interactions, and connectivities from different perspectives than used in the project's case studies. However, a challenge related to working with open data was that datasets are not typically readily accessible or usable to be explored extensively. In the context of future corridor development, hackathons could be organised by using a more focused setting and ready datasets to deliver concrete solutions to specified corridor challenges.

The setting brought together participants with radically different academic disciplines, which comes with its own challenges. However, bringing together radically different disciplines is a necessity when attempting to combine data analysis expertise with the ability to apply the results of technical analysis to wider societal challenges. In the Big Data Hack, participants coming from different fields of research understood data and its purposes very differently, emphasising different aspects of data literacy. While some participants viewed datasets and their varieties as pathways to design concepts encompassing new possible solutions to new problems, others viewed data only as a technical way to addressing already identified problems and needs. In other words, data can be seen in a holistic way as a starting point for identifying previously unknown possibilities or in a more atomistic way as building materials for solutions for a well-defined customer problem.

Relatedly, a hackathon as a type of event is not readily comprehensible for all participants in a similar fashion. While some participants thought the Big Data Hack was a creative event for brainstorming and exploring possibilities of utilizing multiple datasets on a conceptual level, other expected the event to be a competition focused on producing technical ready-made solutions. These characteristics made Big Data Hack both challenging and fruitful. A key takeaway for public organisations is that hackathons can provide both technically-oriented and data-driven solutions for stakeholders, and conceptual models for e.g. generating behavioural insights required for making practical policy choices.

5 IMPLICATIONS FOR CORRIDOR GOVERNANCE: EMPHASIS ON CAPACITY BUILDING

Big data is often mentioned in different public sector strategies and agendas, but the practices of data utilisation are still in their infancy. Often the use of big data in public organisations dealing with territorial development has remained at the level of strategy talks without much concrete development. Therefore, it is important to start capacity building initiatives at both organisational and corridor-wide levels which are grounded in the best practices emerging in companies, research centres, and governments. Proposals for what these kinds of initiatives could be are presented in this chapter. The capacity building aspects are targeted at public sector organisations that are also involved in corridor development, or territorial development more generally.

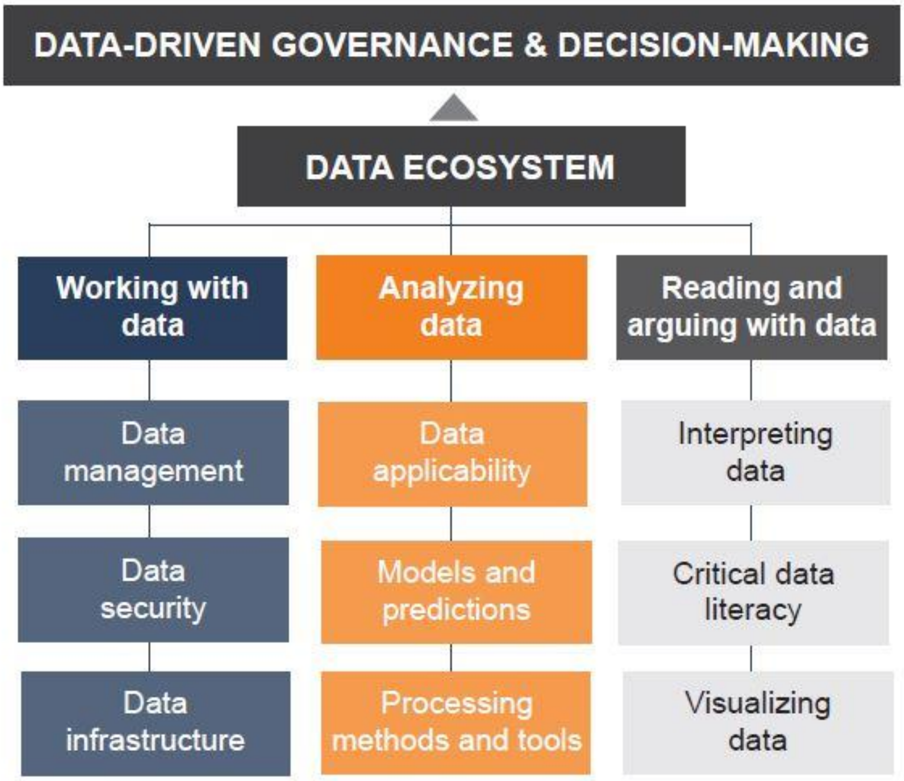
In the next subchapters, recommendations for capacity building to support data-driven corridor governance are first discussed at the level of corridor governance and policy-making, after which different aspects of a thriving data ecosystem are presented. Recommendations for capacity building are loosely following key aspect identified in the research roadmap for Europe to understand different dimensions of digitalisation (Cuquent & Fensel, 2018). All data ecosystem dimensions are presented from the perspective of data literacy emphasizing ability to put data into action (Deahl, 2014). For these purposes, data literacy includes working with, as well as analysing, reading and arguing with data (D'Ignazio & Bhargava 2015).

5.1 FROM DATA MANAGEMENT TO DATA-DRIVEN GOVERNANCE

A more widespread use of evidence in corridor-based policy-making requires a shift in orientation from purely technically-oriented data management to more comprehensive data-driven governance. This shift reflects wider changes that has taken place in territorial governance – namely the shift from centralized and vertical hierarchical forms to decentralized, horizontal and networked forms of governance that support integrated territorial policy-making. In this soft governance environment the role of evidence production and utilisation is also changing. In transnational growth corridors operating in multi-scalar governance environments, the role of strategic evidence is becoming increasingly important for the promotion of corridor objectives. Thus, what counts as evidence depends not only on commonly agreed objectives but also on the wider discussion realm of territorial policy-making. A broadening of the knowledge-base for policy-making is important not only for the legitimation of selected policy paths, but also to challenge the dominant policy knowledges based on a rather one-sided understanding of functionality.

Below critical capacity building aspects for big data driven policy-making are presented in Figure 5.1 and they are explained in detail in the following subchapters. These aspects are discussed from the perspective of how data literacy can be developed in the NGZ. After the presentation of each aspect of the data ecosystem the capacity building testimonials of practitioners in the NGZ are presented. The testimonials are introduced in boxes at the end of each chapter.

Figure 5.1. Recommendations for capacity building in growth corridors. Freely following key aspects identified in the research roadmap for Europe (Cuquent & Fensel 2018).



Capacity building for data-driven corridor governance starts from recognizing the various dimensions of integrated policy development (see Chapter 2). In addition to sectoral and spatially integrative policy-making, combinations of various datasets are needed for more comprehensive understanding about constantly changing territorial dynamics in growth corridors. This sets various requirements for data literacy. Actors working with corridor development should not only understand the holistic nature of corridor development but also know how to select datasets representing different dimensions of corridor development, know how to integrate and analyse them, and also how to combine them with other information sources and prior knowledge (cf. Koltay, 2015). The status of new data sources such as social media in territorial decision and policy-making processes has to be constantly elaborated as insights derived from such data sources are not necessarily easily compatible for example with knowledge based on traditional statistical analyses.

Enormous volumes of data are already generated in public sector organisations, but only some are relevant for corridor level policy-making and most of which describe static and stable conditions. It is important not just to develop data management practices in growth corridors, but also new data-driven governance arrangements need to be established in close collaboration between the public and the private sector. For example, collaboration of public authorities with private companies to gain new insights – e.g. flow analytics services based on mobile positioning data -- can enable speedy utilisation of new data sources.

Altogether, growth corridor governance not only includes public authorities from diverse administrative scales, but also stakeholders such as universities and businesses. Public sector organisations can, for instance, promote open data policies, release open data, and promote common data standards to support interoperability (European Commission, 2014). Universities and other public sector institutions can have a role in R&D to develop reliable algorithms and data analysis tools. Businesses can have a role in developing innovative value-driven applications and solutions, as well as in sharing their data (Cavanillas et al., 2016). At its best, big data can be used to stimulate and inform continuous improvements and align appropriate strategies (Mandinach, 2012).

The ESPON COMPASS project (2018) emphasized that EU institutions and sectoral policies must address their "spatial blindness" and work with planning tools and procedures more effectively. There is a trend towards spatial planning and territorial governance initiatives in functional areas across administrative boundaries such as growth corridors and the trend should be harnessed to secure coordination of the territorial impacts of investments (COMPASS 2018). A range of sophisticated instruments for policy steering – standardization, quality benchmarking and data harmonization – underpins governance changes in corridors. ESPON Cross-border Public Services project also emphasised that the lack of comparability of data is an obstacle for cross-border territorial cooperation. This challenge is also relevant to data-driven governance and policy-making in the growth corridors. Developing a joint harmonised system that takes into account different needs at different territorial scales should be a high-level objective in the EU (CPS 2018).

According to Cavanillas et al. (2016) the dimensions of a coherent big data ecosystem are data availability and access to data; skilled workers; overcoming technical challenges; and an appropriate regulatory environment. In addition, value exists in big data ecosystems in new potentials to transform big data into value-driven applications, new business models, and increased awareness of the benefits of sharing and using big data. When it comes to developing the big data ecosystem of the NGZ, many of these dimensions and values are missing or not fully realised, which points to the need for greater capacity building to create a viable ecosystem.

Box 5.1. Practical views on data-driven governance of the NGZ.

Corridor-level big data strategy

At present, all countries of the NGZ have their own big data strategies, but the corridor itself is lacking one covering its total expanse. The usage of big data could improve data-driven decision-making and leadership at the corridor level. The creation of a strategy for corridor analytics is needed to enhance data-driven policy-making.

Creating the data ecosystem

The level of openness in data policies is dependent on the actions of the actors in the corridor. Public-private partnerships are essential to trigger change. Motivating stakeholders, both public sector actors and private companies to open their data through opportunities and incentives is a key for systemic change. The goal of data-driven governance at the NGZ can be promoted at the ecosystem level. Collaboration and coordination among actors in the corridor are needed to achieve big data-driven governance and evidence-based policy-making.

5.1.1 Working with data

One key aspect of data literacy is be able to work with data (D'Ignazio & Bhargava, 2015), which involves creating and managing data. An aspect of data literacy is *Data management* which refers to, for instance, setting policies and standards for protecting and using data (Tallon, 2013). Nowadays, the lack of standardisation and data management principles make the utilisation of big data challenging. It is possible to standardize data practices by using company-led and industry-led standards (e.g. ISO, MaaS Global API, or Data Catalogue Interoperability Portal standard). More widespread use of standardization practices could facilitate increased big data utilisation in the NGZ. Also new kinds of organization cultures are needed in a data-driven future of growth corridors. For public sector organizations of the NGZ, new types of organisational culture are needed to support knowledge sharing, value analysis, education, cross-department and inter-organizational collaboration, and training, as well as strategic IT planning. At its best, better data management can support practices of good data governance. As leaders prioritise data-informed actions in public sector organisations, new operational models would lead to greater discoverability and usability of data resources.

A second key aspect of working with data is broadening organisational awareness of the need for *data security* and *privacy*. Organisations should consider data protection policies that strike a balance between opportunities and risks. Of these risks, privacy is a critical concern that can limit in what ways big data can be used. Therefore, robust anonymization strategies are important when public sector organisations open data or share data with others. Going further, paying attention to limiting the ways personal information can potentially be de-anonymised is important. Altogether, privacy is a critical factor affecting the ability of governments and other public sector organisations to use big data in territorial development (Stough & McBride, 2014). Still today, very little is understood about the regulatory and ethical implications underpinning the big data phenomenon (Tallon, 2013). Many unconventional data sources have unclear perceptions of data ownership. Regulatory aspects are particularly challenging especially in cross-border territorial development, as organizations might face a complex network of even contradictory country-specific laws. Case study 3 (chapter 4.3) is a good example of this as collecting, storage and processing of the mobile positioning data had to be done in compliance with EU requirements regarding the protection of personal data.

A third element of working with data is *big data infrastructure*, which can include aspects such as building infrastructure for data integration, data governance, data analytics and running APIs (Application Programming Interface). Big data – due to its size and variety – requires greater computing resources. Organisations can gain access to such resources through cloud computing services built for data processing and advanced analysis (e.g. Deep Learning). As more big data is taken into use, organizational data and hardware infrastructure will need to adapt. As infrastructure limitations are reached, organisations will need to find ways they can benefit from advances in extreme computing (e.g. Boman et al. 2015; Feldman 2018).

Box 5.2. Practical views on data management, legislative issues and data infrastructure on the NGZ.

More focus on data management

A lack of data management principles can make the utilisation of big data challenging in growth corridors. It is typical that, e.g., not all departments of the same organization are producing data in consistent formats or quality levels. For instance, metadata describing what the fields contain in a dataset are typically missing. Another challenge is that public sector actors struggle to regularly update the datasets they publish to open data portals which makes these datasets unreliable for others in the data ecosystem. Data management principles focused on providing resources and emphasis on sharing usable and discoverable data at reliable refresh intervals can help address these gaps. As public organisations share and use each other's data in the corridor, the knowledge-base for corridor-level development and decision-making will improve.

Legislation enabling big data utilisation

Legislation that enables big data utilisation is essential for data-driven corridor governance. This legislation should illuminate standards that protect privacy so that some of the most promising data for describing flows and interactions in the NGZ – like social media data and customer data – can be analysed and used to support policy-making. Processing personal data will require new business models as well as scrutinizing data utilisation pathways and data ownership. The regulatory environment should facilitate big data ecosystem by clarifying how data can be shared and exchanged among private and public actors.

Big data infrastructure

Many times, the IT infrastructure of public organisations is not ready for performing advanced analytic operations on big data. For instance, computers and servers may not have enough capacity or be of the proper configurations to run the latest and most advanced machine learning, AI, or deep learning based analytical software. Data infrastructure needs to be improved to meet the needs created by big data, to overcome challenges related to, e.g., large-scale data acquisition, efficient storage and real-time data processing. Furthermore, thinking in terms of corridors, effort should be made to share such resources with all corridor actors – especially between larger cities and smaller rural communities.

5.1.2 Analysing data

In territorial development, data processing and analytics is one of the key areas in need of capacity building. According to D'Igazio & Bhargava (2015) one key data literacy aspect is the ability to analyse data. People working in public organisations need to know at least the basics of data analytics and data science so they can better understand the limitations and benefits of such processes, utilise the insights produced by them in policy-making. Analysing data involves different aspects, such as filtering, sorting, and transforming data. It also involves a larger setting in which data is explored, evaluated, selected, gathered, pre-processed, and then run through an analytical process. The selection process is often nuanced, balancing what data is available with what data is most *applicable* for corridor development and bring value for corridor level decision-making.

An applicable input data can come from different sources. The data, for example, can be open data from public data sources such as governments and agencies, open data portals etc. Examples of non-open datasets are, for instance, customer data, search engine terms, and social media posts. Input data can be both from conventional data sources, like data from population registers or trade statistics; or from unconventional data sources like data from satellites or phone calls. In principle, the resolution of useful input data describing corridor-level interactions can vary from personal, local, municipal, regional to national levels. However, in practice, unconventional datasets which can describe flows and interactions in a new light are often location-based and thus provide more detailed information about corridor dynamics. In addition to different scales, there are also other variables and attributes affecting the quality and applicability of the data, such as level of processing of data and level of structure, refresh frequency and extraction effort of a dataset. (See Chapter 3).

Output data, or processed data, can be used for corridor-level modelling. Examples of output data can be, for instance, *trend predictions and estimations* (e.g. nowcasts) describing interactions and connectivity in the corridor. Processed data can be a combination or restructured version of unconventional input data. For example, unconventional dataset like mobile positioning data enables seeing short-term differences in flows and interactions in Estonia. This can potentially be possible for the whole NGZ, as mobile phone operators have already started to sell their customer data in Finland and Sweden.

Another potential is fundamentally related to data integration, as it is seen that one of the most evident goals of using new unconventional datasets is to generate some new insight not available from conventional data sources alone. All of the case studies of this targeted analysis point to possibilities that exist in combining the new approaches they demonstrate with other conventional forms of data. The temporal and spatial insights that are possible from combining conventional and unconventional datasets can provide significant value for businesses and the public organisations in a growth corridor. Unconventional datasets can be, for instance, social media data or images, photos, videos, space-based monitoring, traffic monitor data, or mobile positioning data. Conventional datasets could include more static population registers, business registers, and economic indicators. When combined with conventional data, for example linked by time stamps, an analysis can reveal insightful temporal patterns. Bringing together conventional and unconventional datasets can shorten the time between documenting observations and producing indicators. Timely insights help policymakers prepare for emerging shocks, stresses, or opportunities across corridors.

In data processing and analytics, suitable methods are needed. The technological capacities regarding big data are rapidly evolving in connection to the development of mainstream and increasingly available artificial intelligence, machine learning, and deep learning tools. Some of these tools are being developed as additions to existing services offered by big technology firms while others are being developed or applied by start-ups. Analytical approaches and tools have are constantly evolving to do more with big data. These tools, when used ethically, can help private and public organisations more quickly make sense of big data and support process improvement, better environmental awareness, and evidence-based decisions.

Appendix 3 presents a few examples of big data tools that are currently available on the market. The presented tools, and others like them, can be useful for processing datasets that are targeted for growth corridors. Some of these tools can potentially support national, regional, and municipal governments and other actors in strategically developing their big data capacities for corridor development. Some have rapid learning curves while others have longer ones. All represent pathways to meaningful skill-building.

All three case studies reveal that new methodologies need to be developed to couple new datasets with appropriate processes to analyse the data. It is evident that methodological development requires time – e.g. Case 1 is at the very early stages of methodological innovation while Case 3 is based on more than a decade of research and development and clearly can go further in what it can achieve because of it. In the development of growth corridors, it is important to give time and resources to support the development of European ecosystems of data management and data analysis methodologies. The EU could support this with targeted capacity building programmes. This ecosystem development has already started as, for example, the European Commission settled an artificial intelligence strategy in 2018 to boost financial support of AI development and to encourage its uptake by public and private sectors (European Commission 2018).

Box 5.3. Practical views regarding development of data quality and data analysis in the NGZ.

Applicability of the data

Many promising sources of big data describing flows and interactions within the corridor are unconventional and more unreliable than the traditional ones such as census data. Therefore, the reliability of new data sources requires scrutiny while the processes to analyse them must be developed in order to produce high-quality and usable data products. Typically, the most reliable data is published by national statistical offices, but these datasets are usually published only once a year and the published data can be as much as two years old. Reducing the time required to produce key indicators and developing new forms of insights can be useful to territorial and growth corridor planning and is an important development issue. Unconventional data sources like social media data or customer data provided by private sector actors hold high potential value from the perspective of spatial and temporal analytics at a corridor-wide scale, but single actors in the corridor do not usually have enough resources to study how to utilise these new data sources while simultaneously taking measures to address their potential biases or other limitation. National statistics offices and the EU should continue taking leading roles in studying how to utilise these new data sources reliable ways with a goal of supporting evidence-based decision-making of growth corridors and territorial development overall.

5.1.3 Reading and arguing with data

Using big data for evidence-based decision-making requires not only technical skills but also critical thinking skills pertinent to data. Data literacy refers not only to the technical or statistical literacy. It also refers to the ability to put data into action (Deahl, 2014). In other words, non-technical actors also benefit from pathways towards data literacy. The *ability to read data* involves the *capacity to interpret data*, for instance, by understanding what data is (and is not), what aspects of the world it represents, and how it impacts people's lives (D'Ignazio & Bhargava, 2015). It is particularly relevant for actors in the NGZ to promote training and professional education for developing data literacy along these lines so that they are better able to recognise the possibilities and limitations of using big data. The NGZ would benefit from having more experts at all levels of governance and activity who have sophisticated and data literacy skills to interpret, apply, and critically contextualize the ways big data can be analysed and used for territorial policy-making.

Critical data literacy affords people the capacity to consider the properties and limits of different datasets, knowledge regarding how to ask questions about how datasets are produced and used, use data to inform decisions, and be more reflective of their personal involvement in data ecosystems. Promoting training and professional education for improving capacities to interpret big data analyses is particularly relevant. In territorial development, policy-makers are often the end users of big data analytics and the focal point of the analytical design processes. For them to utilize big data for supporting corridor objectives, an ability to 'argue with data' is highly relevant convey the meanings of a new insight to larger narratives (D'Ignazio & Bhargava, 2015).

Supporting the ability to 'argue with data' are skills in *visualising big data* analysis results and questioning how others have visualised data. Policy-makers are best able to apply these results when they are in a clear format (e.g. maps, charts, infographics, or animations) ensuring that big data analytics results are understood quickly and easily. To support and legitimise growth corridor development initiatives, visualization of geospatial big data as well as interactions and flows is especially important. New tools help to visualize, for instance, monetary flows and interactions along corridors. Some end user centred analytics tools and visualization applications for geospatial data are already available, but as, for instance, Case 2 results show, more proper visualization approach should be utilised and developed further. Organisations should keep watch for new and innovative ways to visualise data, develop cultures of discussing the merits and limitations of various visualisation outputs, and grow their own capacities making meaningful and information-rich visualisations. Communicating data-driven insights through visualisations in itself can impact growth corridor developments.

Box 5.4. Practical views regarding application and impacts of big data analysis results in the NGZ.

Communicating big data results to decision-makers

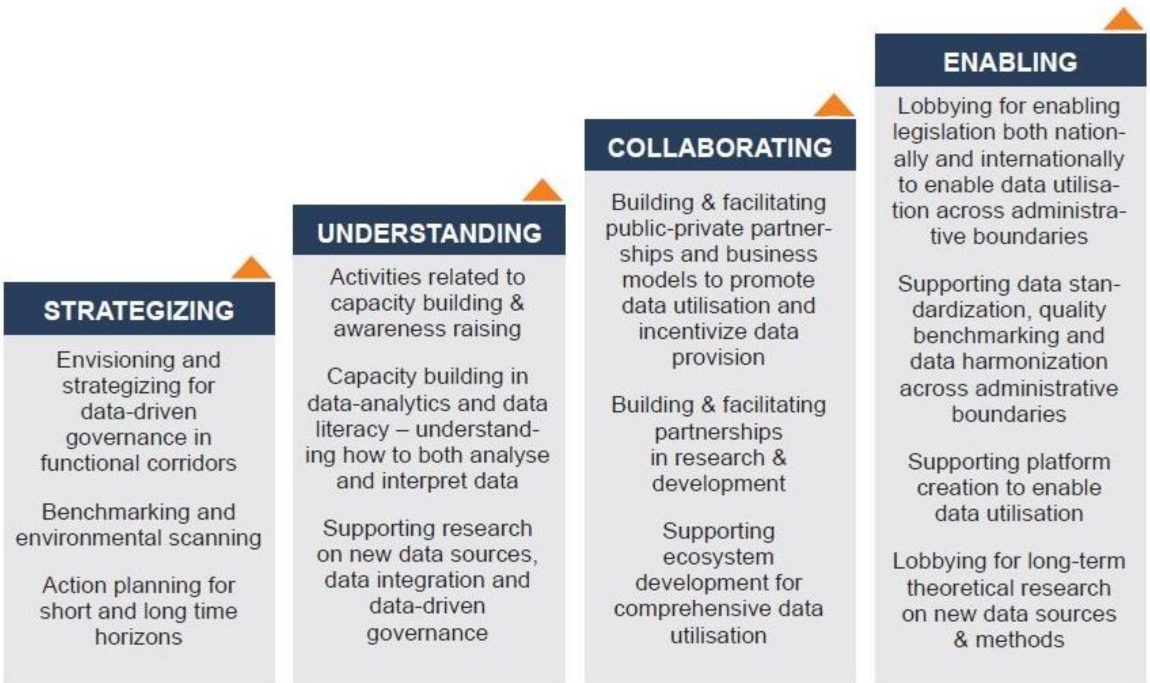
The NGZ data ecosystem needs to develop data literacy of non-technical actors, like experts who are capable of transforming data into appropriable data and concrete action. By this way, big data can be better applied to decision-making and value creation, as data analysts typically are not the right persons to apply data in decision-making. Another data literacy aspect is to develop skills and tools to visualise the results of big data analysis e.g. in the forms of maps and infographics. More emphasis should be placed on capacity building related to spatial analytics especially in the organisations involved in the development of functional regions.

6 SUMMARY OF POLICY RECOMMENDATIONS

Based on the review of the broader big data landscape and the case studies exploring the potentials related to three different datasets made in this targeted analysis, this chapter presents a summary of policy implications as well as recommendations for big data-driven corridor development. The presented recommendations are widely applicable to territorial development as the utilisation of big data is still mostly hindered by general-level challenges. These general challenges include lack of capacities and skills in organisations to utilise new and big datasets, as well as legislative restrictions on wider utilisation.

However, several measures to address these challenges can be promoted through corridor collaboration (Figure 6.1). The measures are further elaborated in the practical guide of this targeted analysis dealing both with collective and individual action that needs to be taken to promote big data utilisation. Investing in exploring the potentials related to big data is a collective effort that needs to be supported by the EU and national-level authorities while engaging multiple stakeholders. At the end, however, the utilisation potential is also relative to the intended use: public-private partnerships with companies offering flow analytics services may provide rapid support for policy-making, whereas harmonisation of statistical systems at the EU level can support integrated territorial development over the long run. Short and long-term objectives are not necessarily mutually exclusive, but there is often a trade-off between resourcing short-term and long-term gains that needs to be considered.

Figure 6.1: Steps towards data-driven corridor governance.



6.1 STRATEGIZING FOR DATA-DRIVEN CORRIDOR GOVERNANCE

Strategizing for data-driven corridor governance is needed to enhance evidence-based policymaking. More comprehensive data utilisation necessitates a clear strategic position when it comes to basing corridor development on an increased understanding of flows and interactions. Here, benchmarking and diverse environmental scanning tools could be utilized to clarify the options. Strategizing also needs to be supported by concrete action planning measures, which have to touch upon the issues presented in the Figure 6.1. Altogether, strategy creation for corridor development could benefit from more comprehensive understanding of corridors as assemblages of physical, social, and digital connectivities, which then creates a good basis for more versatile utilisation of evidence. So far, rather one-sided understandings of corridor functionalities as physical transportation conduits has narrowed what counts as evidence for corridor-wide strategies.

6.2 CAPACITY BUILDING FOR DATA-DRIVEN GOVERNANCE

Using big data for evidence-based decision-making requires both technical skills in organisations and wider data literacy in the broader public. Organisational change starts from highly skilled management with the capability to recognise various data-driven potentials, but at the moment, the NGZ is lacks enough experts capable of transforming their organisations into data-driven ones. Territorial administrations need to improve their readiness and capabilities to recognize potentials in new data sources and generate actionable insights from them. Promoting training and professional education for improving capacities to interpret big data and question insights generated from it is particularly relevant. A broader understanding of how data can be used, the properties and limits of different datasets and knowledge how can inform decisions is needed at all levels of governance. Therefore, all public administrative bodies in growth corridors would benefit from launching initiatives to be fully data-driven in their actions.

6.3 CAPACITY BUILDING IN DATA ANALYTICS

Public organisations should develop their skills in data analytics. One of the most evident additional values from using new datasets is the potential to produce new insights more quickly by combining conventional and unconventional data sources. In addition, integrating physical, social and digital aspects of corridor development can lead to more comprehensive understanding of interactions and flows. For example, the automated traffic measurement data is an example of a dataset that should be used in tandem with other datasets to generate new value for corridor development. A lot depends on the capacities of public organisations to utilise new data sources and develop their skills in data analytics.

However, single actors of the corridor do not have enough resources to study how to utilize new data sources that are less comprehensive and more unreliable that the data provided by statistical offices. Furthermore, the case studies conducted in the project demonstrate some practical challenges linked with utilising big data for corridor-based development policies. Partnerships in research and development – e.g. between public organisations and research organisations or companies – could be established to create public value from new data sources. The EU, as well as national statistical offices, should take a greater role in studying or supporting studies related to reliable and ethical utilisation of new data sources. They should also study or support studies of new analytical tools suitable for integrated approach to territorial development with attention to ethical risks from such things as algorithm bias. As the goal is enriching the evidence base for policy-making, methodological triangulation should also be encouraged to deepen quantitative insights based on big data analyses with qualitative ones.

Organisational capacities in data analytics are also related to open data interfaces, which need to support greater access to spatially and temporally rich datasets with high update frequencies. In addition, visualization of geospatial big data and especially diverse interactions and flows is important to broaden the understanding of corridor functionalities. However, practitioners in the NGZ lack good visualisation tools and skills suitable for new available datasets. To be able to use big data for boosting growth corridor competitiveness, potential big data visualization tools and applications need to be assessed and developed, as well as skills to use these new tools. Here, collaboration between public and private sector could be promoted.

The experiments conducted in the Case 2 pointed out that big data analysis of new datasets can reveal potentials for more integrated corridor development. Network analysis on diverse interactions should be utilised and developed further for example for a large-scale trend analysis concerning the development of European growth corridors. In the NGZ, the results could be used to support networking especially within the western parts of the corridor especially in Sweden and Norway, and among all nations along the corridor.

The potentials in using mobile positioning data relate to almost real-time understanding of mobility dynamics that cannot be captured by using traditional datasets. The utilisation of this data should be promoted and combined with the existing knowledge base on mobility. As a result of the work conducted in the Case 3, the resulting database supports mobility-related policy-making in Estonia. The lessons from the Estonian case should be utilized in diverse European countries aiming to promote the utilisation of similar datasets.

6.4 HARMONIZED DATA MANAGEMENT SYSTEM

The lack of clear and harmonized data management principles makes the utilisation of big data challenging in growth corridors, especially when crossing national boundaries. It can be an obstacle in cross-border cooperation if the data provided by different actors and authorities is not comparable or the data has diverse quality problems. Therefore, integrating various data sources necessitates solid structures for integrated data management. Developing a harmonized system that takes into account special needs of corridor governance should be a common objective. Data standardization, quality benchmarking and data harmonization are essential goals for such efforts. The EU can support European ecosystems of data management with capacity building programmes.

Furthermore, long-term access to data should be secured, as for example, it is now difficult to have confidence that datasets on open data portals will be updated consistently over time. Data portals should be updated regularly so that they can be used as reliable data sources for corridor development and decision-making. Data from private-sector players also introduces potential surprises as companies can go bankrupt, merge, or arbitrarily change their policies or prices regarding the data they provide. As value is coupled to various data sources, alternative long-term storage and access solutions will become increasingly necessary.

6.5 PUBLIC-PRIVATE PARTNERSHIPS TO FOSTER BIG DATA UTILISATION

Realizing the potentials of big data requires partnerships among public authorities, private providers, and research organisations within an overall ecosystem fostering sophisticated utilisation of big data. New collaborative business models should be developed to incentivise data provision, as there are relatively long value chains related for example to mobile positioning data requiring expertise from several research fields. Thus, the wider utilisation of mobile positioning data would clearly benefit from public-private partnerships, as the data also allows the estimation of flows for cross-border movements based on mobile network operators roaming data. Methodologies can be shared to develop transnational mobility insights that paint an even more comprehensive pictures of growth corridors. Partnerships between public authorities and private companies at diverse scales of development can aid achieving such aims. Furthermore, strengthening digital infrastructures in both urban and rural areas along the growth corridors is necessary to even out potential downstream disadvantages to rural areas as access to data-driven insights are applied in governance processes.

6.6 COLLABORATION IN RESEARCH AND DEVELOPMENT

Further research is still needed on how to foster big data utilisation in territorial development and corridor-based development policies as well as on revealing territorial dynamics in functional areas. Such data-driven governance of growth corridors can challenge and renew the spatial organization of societies based on increased understanding about contemporary spatial connectivities. Contemporary understanding of functional corridors needs to be challenged by incorporating diverse spatial understandings through a versatile use of data to corridor analytics.

Furthermore, the case studies conducted in the project demonstrate some practical challenges linked with utilising big data for corridor-based development policies, such as the private ownership of some of key datasets describing flows and interactions. Therefore, partnerships in research and development – e.g. between research organisations and companies – could be encouraged to create public value from new data sources. Also, ethical issues and privacy concerns, and the lack of ground truth data to validate some findings may hinder the utilisation of new datasets. New methodologies and models thus need to be developed to couple new datasets with advanced processes to analyse the data. This often requires long-term theoretical study, which should be supported through research funding mechanisms.

For example, automated traffic measurement data has many good attributes that make it interesting in terms of research, practical flow tools, and analytics. The measurement system has been online a long while and there is now data available spanning multiple years. The network of stations generating data is still sparse on some roads, but for E18 and the NGZ the dataset is rich and enables analytical tools to be applied to gain more insight to policy-making. Predictive models are among the toolsets that policy makers can develop together with researchers to leverage this dataset for many applications.

Also, collaboration between a ministry and a university in the case 3 is a good example of collaboration in research and development to promote the utilisation of new data sources. In the case study, the Mobility Lab of the University of Tartu developed and implemented a methodology for everyday mobility database that is now set to become an important input for the ministry's spatial planning and decision-making processes to answer questions related to transportation and mobility.

6.7 PLATFORMS TO ENABLE WIDESPREAD DATA UTILISATION

The realisation of a data-driven strategy requires concrete initiatives to develop collaborative platforms for corridor governance to create public value from new data sources. This, however, necessitates a change in thinking when it comes to territorial governance as corridors are increasingly seen as comprehensive governance frameworks for wider societal change based on functional geographies. For example, the promotion of fast train connections in the study area of this targeted analysis is a clear example of a dominant spatial development logic leaning increasingly on functional geography and accessibility in relation to improved transportation networks. From this perspective — if functional cross-border geographies are increasingly seen as a basis for spatial governance arrangements — big data has a clear potential in renewing or even disrupting transnational corridor governance based on new collaborative production of data-driven insights and new business models between public and private actors. However, the democratic legitimisation of these new soft governance frameworks needs to be ensured, for example, by innovating new forms of participatory democracy.

6.8 LEGISLATION ENABLING DATA-DRIVEN GOVERNANCE

Privacy concerns, obscurity of data ownership and ethical issues may hinder the utilisation of new datasets for policy-making. Legal clarity is needed to enable big data utilisation for corridor governance, especially in cross-border areas. In addition, privacy aspects should be studied further as some datasets describing flows and interactions like social media data and customer data legitimately raise privacy concerns that must be balanced with the potential public value of use-cases for such data in policy-making. Many of these concerns are merited as de-anonymization becomes ever more possible when technology enables more datasets to be efficiently combined by interested parties to reassemble obscured identities. To promote wider utilisation of mobile positioning data, for example, further attention needs to be directed to the relationship between mobile positioning data and EU legislation.

7 CONCLUSIONS

Several ongoing trends such as digitalisation, rapid development of new data analytic tools, and increased opening and sharing of data combine to make big data a significant source of new insights for territorial development. Especially the rapid increase in new data sources related for example to sensor and customer data is seen to be changing the scheme of spatial analytics. The objective of the Big Data & EGC targeted analysis project was to generate understanding about the various potentials that big data could bring to the field of territorial development. The focus was particularly on European growth corridors due to the existing knowledge gaps concerning corridor functionalities; most of the datasets used in corridor development today still represent rather conventional ways of approaching territorial development from the perspective of an administratively and territorially bounded rationality. In addition to this territorially bounded understanding, the analysis of European growth corridors necessitates relational understanding of spatial connectivities that can significantly enrich current policy-making processes.

In the context of European growth corridors big data has already proven to hold potential as, for example, novel concepts supporting sustainable mobility (e.g. Mobility as a Service) necessarily rely on the utilisation of big and real-time data. However, broadening the perspective is still needed to increase the understanding about constantly evolving territorial dynamics that challenge the role of various administrative boundaries, and to increase the legitimacy of the corridor-based development policies. For example, the weak connectedness of regions revealed in Case 2 exploring project partnership data questions the existence of a transnational growth corridor, but the situation might look very different when analysed from the perspective of physical transportation flows. Therefore, corridor functionalities have to be analysed from different perspectives as well as elaborated in terms of their interrelatedness.

This targeted analysis took a step towards this direction and provides a good basis for future research and development to broaden the understanding of functional growth corridors. Collective efforts should be made for example by the EU and statistical offices to further explore the utilisation potential of big data together with research organisations. Altogether, the presented conceptual approach in the project can be utilized to widen the horizon of big data utilisation as it sheds light on the three different yet inherently interconnected physical, social and digital dimensions of corridor development. In addition, the conceptual framework can be used as a practical tool in selecting datasets from different corridor development perspectives. The three case studies conducted in the project demonstrate the complex nature of corridor development as for example data related to mobile positioning can be used in numerous contexts ranging from planning cross-border services to enrich understanding individual activities within corridors.

In fact, portraying European growth corridors as frameworks for meta-governing spatial development highlights the role of evidence – and data – in their planning processes. Data-driven insights are highly functional communication and rallying mechanism in this setting as corridors don't constitute a legally binding framework for territorial development. Rather, corridors, when supported by evidence-base demonstrating the existence of various functional geographies, act as a coordinative and integrative governance framework for numerous co-existing spatialities. Here, big data plays a central role in understanding the complex, often place-based yet spatially interconnected development challenges, and spatial connectivities that cannot be captured by using traditional statistical datasets. The existing gaps in the knowledge base could be significantly complemented by utilising big data sources and new methods of analysis, as well as methodological triangulation combining quantitative and qualitative methods.

Potentials of big data for integrated territorial policy development relate closely to the notion of an integrated approach to territorial development. Corridors per se function as a framework for sectorally and spatially integrative policy-making. Yet, when looking at the potentials of big data for corridor development, also a temporal perspective has to be added to the notion of an integrated approach due to the potentials that new datasets bring to decision-making based on (almost) real-time data. In addition, an integrated approach is fundamentally related to combining datasets, as doing so can produce new insights.

Integrating various data sources necessitates dynamic structures for data management. Integrative and collaborative platforms should be developed for data-driven corridor governance to create public value from new data sources. From this perspective, big data has a clear potential in renewing or even disrupting cross-border corridor governance based on new collaborative business models between public and private actors, if new knowledge about functional geographies is used to challenge traditional territorially bounded understanding. A lot depends on the capacities of public organisations to utilize new data sources and develop their skills in data analytics, which requires updated expertise in data-driven governance. In addition to supporting evidence production, support has to be given also to capacity building related to the use of this evidence. Leadership is needed to promote data-utilisation throughout public organisations. In collaborative platforms for more enhanced data utilisation, the added value of data integration is driven primarily by an improved understanding of various co-existing spatial connectivities that can be utilized in parallel to promote place-sensitive policies.

Yet, many challenges in utilising big data in corridor development need to be taken into account, such as the missing spatial dimension from big data components. In addition, sharing data and information between private providers, authorities and countries is a special challenge that needs to be addressed in the forthcoming collective attempts to foster big data utilisation. In fact, private companies own some of the most promising dataset describing corridor flows and interactions, thus creating a challenge of data accessibility. Also, ethical issues and privacy concerns as well as a lack of ground truth data to validate the findings may hinder the utilisation of new datasets. Furthermore, new methodologies need to be developed to couple new datasets with new processes for analysis via research collaborations and sufficient support for basic research. Overall, as big data utilisation comes with a cost, careful attention should be paid to the measurement of interactions and flows through new data sources and thus institutionalising new data analysis practices.

In the context of European growth corridors, potentials of big data also relate to broader discussions about the value-base of European territorial development. A risk of a digital divide has been identified that could further widen the gap between urban and rural areas when it comes to resources and readiness in data utilisation. Corridor policies could either enable widening or bridging the gap between urban and rural areas depending on the broader societal context and value-base of territorial development. Here, good practices related to data utilisation in corridor development should be made mainstream to allow territorial integration. For example, the experiences in Estonia could be followed to deepen the understanding of how to use new data sources in territorial policy-making. At best, big data can increase the understanding of place-specific particularities and functional geographies between dense and sparse areas that could be used for the benefit of lagging areas. Here, linking the technical results of new big data analyses to bigger territorial narratives grounded in a common value-base is important for more efficient utilisation of spatial analytics and thus for mobilising the potentials related to territorial integration. Narratives about new functional geographies should be directed not only to European policy-makers but also to European citizens who could also benefit from seeing their communities and daily geographies in a new light. In conclusion, the potentials of big data for integrated territorial policy relate first and foremost to seeing things differently: big data can act as an eye-opener about new spatial realities that challenge the current ways of organising societal practices and territorial governance, both in European growth corridors and territorial development more generally.

REFERENCES

Ahas, R., Aasa, A., Silm, S., Aunap, R., Kalle, H. & Mark, Ü. (2007) Mobile Positioning in Space–Time Behaviour Studies: Social Positioning Method Experiments in Estonia. *Cartogr. Geogr. Inf. Science,* Vol. 34: 259–273.

Ahas, R., Aasa, A., Roose, A., Mark, Ü. & Silm, S. (2008) Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, 29(3): 469–486.

Ahas, R., Aasa, A., Silm, S. & Tiru, M. (2010a) Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area : Case study with mobile positioning data. Transp. Res. Part C 18, 45–54. https://doi.org/10.1016/j.trc.2009.04.011

Ahas, R**.**, Silm, S., Järv, O., Saluveer E., Tiru, M. (2010b) Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones, *Journal of Urban Technology*, 17(1): pp. 3–27.

Ahas, R., Aasa, A., Yuan, Y., Raubal, M., Smoreda, Z., Liu, Y., Ziemlicki, C., Tiru, M., Zook, M., 2015. Everyday space – time geographies : using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn. *Int. J. Geogr. Inf. Sci.*, Vol. 29: 2017–2039. https://doi.org/10.1080/13658816.2015.1063151

Ahn, Jae-wook, Meirav Taieb-Maimon, Awalin Sopan, Catherine Plaisant, and Ben Shneiderman, "Temporal Visualization of Social Network Dynamics: Prototypes for Nation of Neighbors", in John Salerno, Shanchieh Jay Yang, Dana Nau, and Sun-Ki Chai (eds.), Social Computing, Behavioral-Cultural Modeling and Prediction, Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2011, pp. 309–316.

Ahuja, R.K; Magnanti, T.L. & Orlin, J.B. (1993) Network Flows: Theory, Algorithms, and Applications. Pearson Education Inc.

Ambrosino, G., Nelson, J. D., Boero, M., & Pettinelli, I. (2016) Enabling intermodal urban transport through complementary services: From Flexible Mobility Services to the Shared Use Mobility Agency: Workshop 4. Developing inter-modal transport systems. *Research in Transportation Economics*, Vol. *59*, 179–184. https://doi.org/10.1016/J.RETREC.2016.07.015

Astrov, V., Grieveson, R., Hanzl-Weiss, D., Hunya, G., Jestl, S. & Mara, I. (2018) Socio-economic challenges, potentials and impacts of transnational cooperation in central Europe. Interreg Central Europe, Final report. The Vienna Institute for International Economic Studies. Available at: https://www.interreg-central.eu/Content.Node/events/Final-Report-181018-full.pdf (Accessed 16 October 2018).

Athey S (2017). Beyond prediction: Using big data for policy problems. Science 355, 6324, 483-485.

Bera, Sharminda, and K. V. Krishna Rao, "Estimation of Origin-Destination Matrix from Traffic Counts: The State of the Art", *European Transport \ Trasporti Europei*, No. 49, 2011, pp. 2–23, available at: https://ideas.repec.org/a/sot/journl/y2011i49p2-23.html.

Blomqvist, Kirsimarja, and Juha Levy, "Collaboration Capability a Focal Concept in Knowledge Creation and Collaborative Innovation in Networks", *International Journal of Management Concepts and Philosophy*, Vol. 2, No. 1, 2006, pp. 31–48.

Blumenstock, J. and Fratamico, L. (2013) 'Social and spatial ethnic segregation: a framework for analyzing segregation with large-scale spatial network data', paper presented at the 4th Annual ACM Symposium on Computing for Development (DEV '13), Cape Town, 6–7 December [WWW document]. URL http://www.jblumenstock.com/files/papers/jblumenstock_dev2013.pdf (Accessed 29 September 2016).

Boman, E., Madduri, K., Rajamanickam S & Wolf, M. (2015) High-Performance Computing for Extreme-Scale Data Analytics. US Department of Energy's Office of Scientific and Technical Information. Available at: https://www.osti.gov/biblio/1253084 (Accessed 4 October 2018.)

Boyandin, I., Bertini, E., Bak, P., & Lalanne, D. 2011. Flowstrates: An approach for visual exploration of temporal origin-destination data. *Computer Graphics Forum*, Vol. 30(3): 971-980.

Bramoullé, Y., Galeotti, A. & Rogers, B. (2016) *The Oxford Handbook of the Economics of Networks*. Oxford, UK: Oxford University Press.

Briscoe G. & Mulligan C. (2014) *Digital Innovation: The Hackathon Phenomenon*. CREATIVEWORKS LONDON WORKING PAPER NO.6

Broad, W. J. (2018) It's a Briefcase! It's a Pizza Box! No, It's a Mini Satellite: Orbiting instruments are now so small they can be launched by the dozens, and even high school students can build them. The New York Times, 11 December 2018. (Accessed 31 January 2019.) Available at: https://www.nytimes.com/2018/12/11/science/satellites-space-nasa.html.

Calabrese, F., Pereira, F. C., Di Lorenzo, G., Liu, L. and Ratti, C. (2010) The geography of taste: analyzing cellphone mobility and social events. *Pervasive Computing LNCS,* 6030: 22–37.

Caprotti F., Cowley, R., Datta, A., Castan Broto V., Gao, E., Georgeson L., Herrick C., Odendaal N., & Joss S. (2017). The New Urban Agenda: key opportunities and challenges for policy and practice. *Urban Research & Practice*, Vol. 10(3): 367-378.

Cavanillas,J.M.; Curry, E. & Wahlster, W. (2016) The Big Data Value Opportunities. In: *New Horizons for a Data-Driven Economy. A Roadmap for Usage and Exploitation of Big Data.* (eds.) Cavanillas et al. Springer, Switzerland.

Cavanillas,J.M.; Curry, E. & Wahlster, W. (2016; Eds) *New Horizons for a Data-Driven Economy. A Roadmap for Usage and Exploitation of Big Data.* Springer, Switzerland.

CEC (Commission of the European Communities) (1999) European Spatial Development Perspective: Towards Balanced and Sustainable Development of the Territory of the EU (Luxembourg, Office for Official Publications of the European Communities).

Chen Y-C & Lee J (2018) Collaborative data networks for public service: governance, management, and performance. Public Management Review 20: 5, 672–690.

Chen, X.; Li, L. & Shi, Q. (2014) Stochastic Evolutions of Dynamic Traffic Flow, Springer.

Chui, M., Manyika, J., Miremadi, M., Henke, N., Chung, R., Nel, P., & Malhotra, S. (2018) Notes from the AI frontier. Insights from hundreds of use cases. Discussion Paper. McKinsey Global Institute. April 2018. Available at: https://www.mckinsey.com/mgi/ (Accessed 27 September 2018.)

Combes, P.-P., Overman, H. G. (2003) The spatial distribution of economic activities in the European Union. CEPR Discussion Paper No. 3999.

Crawford, K. (2013). The hidden biases of big data. Harvard Business Review Blog, 1 April 2013. Available at: https://hbr.org/2013/04/the-hidden-biases-in-big-data (Accessed 14 September 2018.)

COMPASS (2018). COMPASS - Comparative Analysis of Territorial Governance and Spatial Planning Systems in Europe. Final Report. ESPON, 2018.

CPS (2018). Cross-border Public Services (CPS). Targeted Analysis. Draft Final Report. ESPON, 2018.

Cuquet, M. & Fensel, A. (2018) The societal impact of big data: A research roadmap for Europe. *Technology in Society,* Vol. 54: 74-86.

Davoudi, S. (2008) Conceptions of the city-region: a critical review. Proceedings of the Institution of Civil Engineers - Urban Design and Planning 161, 2.

Davoudi, S. (2009) Asymmetric development in spatial planning: Positivist content and post-modernist processes?, in: S. Davoudi & I. Strange (Eds) Conceptions of Space and Place in Strategic Spatial Planning, pp. 207–243 (New York, Routledge).

Deahl, Erica Sachiyo, "Better the Data You Know: Developing Youth Data Literacy in Schools and Informal Learning Environments", Master of Science, Massachusetts Institute of Technology, 2014. Available at: https://cmsw.mit.edu/wp/wp-content/uploads/2016/06/233823808-Erica-Deahl-Better-the-Data-You-Know-Developing-Youth-Data-Literacy-in-Schools-and-Informal-Learning-Environments.pdf.

Dean, Jeff (2019) Looking Back at Google's Research Efforts in 2018. Google AI Blog, 15 January 2019. Available at: https://ai.googleblog.com/2019/01/looking-back-at-googles-research.html (Accessed 31 January 2019).

De Landa, M. (2006) *A new philosophy of society: assemblage theory and social complexity*. Continuum, London.

Deng, Q. & Cheng, L. (2013) Research Review of Origin-destination Trip Demand Estimation for Subnetwork Analysis. In: Procedia – Social and Behavioral Sciences, Vol. 96, November 2013, pp. 1485–1493.

Diebold, F.X. (2000) Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting," Discussion Read to the Eighth World Congress of the Econometric Society, Seattle, August. Available at: http://www.ssc.upenn.edu/~fdiebold/papers/paper40/temp-wc.PDF (Accessed 6 February 2019.)

D'Ignazio C. & Bhargava R. (2015) Approaches to Building Big Data Literacy. Proceedings of the Bloomberg Data for Good Exchange Conference. 2015.

Drewello & Scholl (2015) Integrated spatial and transport infrastructure development: The case of the European North-South corridor. Cham: Springer.

Dühr S, Colomb C & Nadin V 2010. European spatial planning and territorial cooperation. Routledge, New York.

Enoch M 2012. Sustainable Transport, Mobility Management and Travel Plans. Routledge, London.

Erhardt, G. D.; Batty, M. & Arcaute, E. (2018) Recommendations for big data programs at transportation agencies. In: Schintler, L. A., & Chen, Z. (eds.) *Big Data for Regional Science*. Oxon: Routledge.

European Commission (2014) *Towards a thriving data-driven economy. Communication from the commission to the European Parliament, the council, the European economic and social Committee and the committee of the regions,* Brussels.

European Commission (2018) Artificial intelligence: Commission outlines a European approach to boost investment and set ethical guidelines. European Commission press release, 25 April 2018. Available at: http://europa.eu/rapid/press-release_IP-18-3362_en.htm (Accessed 7 February 2019).

European Commission (2019) European Commission's Regional and Urban Policy vision. Available at: https://ec.europa.eu/info/departments/regional-and-urban-policy/mission-statement-regional-and-urban-policy_en

GDPR Section 5 Article 40 REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679#d1e3885-1-1

Faludi, Andreas (2010) Cohesion, coherence, cooperation: European spatial planning coming of age? Routledge, London.

Faludi, Andreas (2018) The Poverty of Territorialism. A Neo-Medieval View of Europe and European Planning. Edward Elgar, Cheltenham.

Feldman, Michael (2018) 'Europeans Budget 1.4 Billion Euros to Build Next-Generation Supercomputers'. *Top 500 Supercomputer Sites,* 1 October 2018. Available at: https://www.top500.org/news/europeans-budget-14-billion-euros-to-build-next-generation-supercomputers/ (Accessed 4 October 2018.)

Feldman, Michael (2018b) IBM Introduces Software to Alleviate AI Bias. *Top 500 Supercomputer Sites,* 24 September 2018. Available at: https://www.top500.org/news/ibm-introduces-software-to-alleviate-ai-bias/ (Accessed 5 October 2018.)

Foray, D., Goddard, J., Beldarrain, X. G., Landabaso, M., McCann, P., Morgan, K., Nauwelaers, C., & Ortega-Argilés, R. (2012) Guide to research and innovation strategies for smart specialisation (RIS 3). Available at: http://s3platform.jrc.ec.europa.eu/c/document_library/get_file?uuid=a39fd20b- 9fbc-402b-be8c-b51d03450946&groupId=10157. (Accessed 16 October 2018.)

Gianelle, C. & Kleibrink, A. (2015) Monitoring Mechanisms for Smart Specialisation for Smart Specialisation Strategies. JRC Technical Reports. S3 Policy Brief Series 2015. Available at: http://s3platform.jrc.ec.europa.eu/documents/20182/114948/JRC95458_Monitoring_Mechanisms_S3_Policy_Brief.pdf/ce74fd68-cd17-4574-950d-4551582655d2 (Accessed 10 September 2018.)

Gianelle, C., Guzzo, F., Mieszkowski, K. (2017) Smart Specialisation at work: Analysis of the calls launched under ERDF Operational Programmes. JRC Technical Reports. S3 Working Paper Series No. 11/2017. Available at: http://s3platform.jrc.ec.europa.eu/documents/20182/222593/ Smart+Specialisation+at+work_Analysis+of+the+calls+launched+under+ERDF+Operational+Programmes/ed9f9e8d-c2e1-4667-832f-f1551f62f583 (Accessed 10 September 2018.)

Grinberger, A.Y. & Felsenstein, D. (2018) Using big (synthetic) data to identify local housing market attributes. In: Schintler, L. A., & Chen, Z. (eds.) *Big Data for Regional Science*. Oxon: Routledge.

Guo, D. (2009) Flow Mapping and Multivariate Visualization of Large Spatial Interaction Data. IEEE Transactions on visualization and computer graphics, 15(6), 1041-1048.

Gutiérrez J.; Condeço-Melhorado A.; López, E. & Monzón, A. (2011) Evaluating the European added value of TEN-T projects: a methodological proposal based on spatial spillovers, accessibility and GIS. *Journal of Transport Geography*, Vol. 19: 840–850.

Hargittai, E. (2015) 'Is bigger always better? Potential biases of big data derived from social network sites.' *The ANNALS of the American Academy of Political and Social Science*, Vol. 659(1): 63-76.

Harris I, Wang Y & Wang H 2015. ICT in multimodal transport and technological trends: Unleashing potential for the future. International Journal of Production Economics Vol. 159, 88-103.

Harjunen, O., Saarimaa, T., Tukiainen, J. (2017) Political Representation and Effects of Municipal Mergers. VATT Institute for Economic Research. VATT working papers 98. Available at: https://vatt.fi/artikkeli/-/asset_publisher/political-representation-and-effects-of-municipal-mergers (Accessed 3 October 2018.)

Harvey, D. (1973) *Social justice and the city.* Edward Arnold, London.

Heikkilä, S. (2014) Mobility as a Service: A Proposal for Action for the Public Administration, Case Helsinki. MS thesis. Aalto University, Espoo, Finland. Available at http://urn.fi/URN:NBN:fi:aalto-201405221895.

Huggins, R. & Thompson, P. (2017) Networks and regional economic growth: A spatial analysis of knowledge ties. In: Environment and Planning A: Economy and Space. Vol. 49 (6), pp. 1247-1265.

Hwang, S., Yalla, S. & Crews, R. (2018) Processing uncertain GPS trajectory data for assessing the locations of physical activity. In: Schintler, L. A., & Chen, Z. (eds.) *Big Data for Regional Science*. Oxon: Routledge.

IBM (2018) Maersk and IBM Introduce TradeLens Blockchain Shipping Solution. IBM News Room, 9 August 2018. https://newsroom.ibm.com/2018-08-09-Maersk-and-IBM-Introduce-TradeLens-Blockchain-Shipping-Solution (Accessed 1 February 2019).

Isaacman, S., Becker, R., Caceres, R., Kobourov, S., Martonosi, M. and Rowland, J. (2011) Ranges of human mobility in Los Angeles and New York, paper presented at the 8th International Workshop on Managing Ubiquitous Communications and Services (MUCS), Seattle, 21 March 2011. Available at: https://raptor.cs.arizona.edu/~kobourov/mucs.pdf (Accessed 7 February 2019).

Jauhiainen JS & Moilanen H 2011. Towards fluid territories in European spatial development: regional development zones in Finland. *Environment and Planning C: Government and policy,* Vol. 29: 728−744.

Jessop, B. (2006) Spatial fixes, temporal fixes, and spatio-temporal fixes. In N. Castree and D. Gregory, eds., *David Harvey: a Critical Reader*, Oxford: Blackwell, 142-66, 2006.

Jones M. (2009) Phase space: geography, relational thinking, and beyond. *Progress in Human Geography*, Vol. 33(4): 487–506.

Järv, O., Ahas, R., Witlox, F., (2014). Understanding monthly variability in human activity spaces : A twelve-month study using mobile phone call detail records. Transp. Res. Part C, Vol. 38: 122–135. https://doi.org/10.1016/j.trc.2013.11.003

Järv, O., Tenkanen, H., Salonen, M., Ahas, R., & Toivonen, T. (2018). Dynamic cities: Location-based accessibility modelling as a function of time. *Applied Geography*, *95*(February), 101–110. https://doi.org/10.1016/j.apgeog.2018.04.009

Kalliomäki H (2012). Towards Comprehensive Spatial Development in Europe: A Critical View from Finland. *Planning Theory and Practice*, 13:4, 569–589.

Kalliomäki H, Saarni J, Sainio J, Vainio J, Antikainen J, Heikkinen B, Laasonen V, Huttunen J, Vähäjylkkä K, Tolonen S & Raninen M 2018. Meriklusterin verkostoanalyysi Pohjoisella kasvuvyöhykkeellä (Translation: Network analysis of maritime cluster in the Northern Growth Zone). *Uudenmaan liiton julkaisuja*, C 86.

Kamel I, Abdelgawad H & Abdulhai B 2016. Transportation Big Data Simulation Platform for the Greater Toronto Area (GTA). SmartCity 360°, 443–454.

Kim, G.-H., Trimi, S. & Chung, J.-H. (2014) Big-data applications in the government sector. *Communications of the ACM.* Vol. 57:No.3 March 2014.

Kanellos, Michael (2016) The Five Different Types of Big Data. *Forbes.com* 11 March 2016 (Accessed 10 September 2018) https://www.forbes.com/sites/michaelkanellos/2016/03/11/the-five-different-types-of-big-data/

Kelly, G. (2018) Apple working on long-range '3D Camera' for 2019 iPhone. *Forbes.com* Available at: https://www.forbes.com/sites/gordonkelly/2018/12/30/apple-iphone-xi-11-xs-max-xr-upgrade-price-cost-3d-camera-iphone-11/ (Accessed 24.01.2019).

Kloosterman, R. C. & Musterd, S. (2001) The Polycentric Urban Region: Towards a Research Agenda. *Urban Studies*, Vol. 38(4), 623–633.

Koltay, T. (2015) Data literacy: in search of a name and identity. *Journal of Documentation*, 71: 2, 401–415, https://doi.org/10.1108/JD-02-2014-0026.

Kotilainen A, Santala V, Helminen V, Kalliomäki H, Sainio J & Koponen A 2016. Esitutkimus: Tietoon perustuva päätöksenteko Pohjoisella kasvuvyöhykkeellä (In English: Knowledge-based decision making in the Northern Growth Zone). *Turun yliopiston kauppakorkeakoulu, CCR tutkimuspalvelut, Sarja* E1/2016. http://www.doria.fi/handle/10024/127272

Kuhn, Max & Johnson, Kjell (2013). *Applied Predictive Modeling*. New York, NY: Springer New York.

Kuhn M 2015. Peripheralization: Theoretical Concepts Explaining Socio-Spatial Inequalities, European Planning Studies, 23:2, 367–378.

Laney, D. (2001) 3-D Data Management: Controlling Data Volume, Velocity and Variety, META Group Research Note, February 6. Available at: https://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/ (Accessed 7 February 2019.)

Levine, Ben, “Pittsburgh Uses Data to Predict Fire Risk”, *Government Technology*, 25 January 2018. https://www.govtech.com/data/Pittsburgh-Uses-Data-to-Predict-Fire-Risk.html. (Accessed 27 June 2019)

Liu, N.X., Baras, J.S. (2003) Statistical modeling and performance analysis of multi-scale traffic. In: IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies, San Francisco, CA, Vol. 3: 1837–1847.

Louail, T., Lenormand, M., Cantu Ros, O.G., Picornell, M., Herranz, R., Frias-Martinez, E., Ramasco, J.J., Barthelemy, M., 2015. From mobile phone data to the spatial structure of cities. Sci. Rep. 4, 5276. https://doi.org/10.1038/srep05276

Malokin, Aliaksandr, Giovanni Circella, and Patricia L. Mokhtarian (2015), "How Do Activities Conducted While Commuting Influence Mode Choice? Using Revealed Preference Models to Inform Public Transportation Advantage and Autonomous Vehicle Scenarios", *Transportation Research Part A: Policy and Practice*, Vol. 124, June 1, 2019, pp. 82–114. https://doi.org/10.1016/j.tra.2018.12.015

Mandinach, E. (2012) 'A perfect time for data use: Using data driven decision making to inform practice'. *Educational Psychologist* 47(2): 71–85.

Marbán, Ó., Mariscal, G. Segovia, J. (2009); *A Data Mining & Knowledge Discovery Process Model*. In Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and Adem Karahoca, ISBN 978-3-902613-53-0, pp. 438–453, February 2009, I-Tech, Vienna, Austria.

Mariussen, Å., Rakhmatullin, R., Stanionyte, L. (2016) *Thematic Work on the Understanding of Transnational cooperation and Value Chains in the context of Smart Specialisation. Smart Specialisation: Creating Growth through Transnational cooperation and Value Chains.* JRC Science for Policy Report. Available at: http://s3platform.jrc.ec.europa.eu/documents/20182/154989/ Smart+Specialisation+Creating+Growth+through+Trans-national+cooperation+and+Value+Chains/f3587cf2-c511-42a4-b559-a89d511a7e38 (Accessed 3 October 2018.)

Mavrič, J. & Bobek, V. (2015) 'Measuring Urban Development and City Performance', In: Perspectives on Business and Management, Vito Bobek, IntechOpen, 2015.

Medeiros, E. (2019; Ed). *Territorial Cohesion: The Urban Dimension*. Springer.

Meijers E. (2005) Polycentric urban regions and the quest for synergy: Is a network of cities more than the sum of the parts? Urban Studies (Routledge). April 2005, Vol. 42(4): 765–781.

Merino, L., Seliner, D., Ghafari, M., Nierstrasz, O. (2016) CommunityExplorer: A Framework for Visualizing Collaboration Networks. IWST'16 August 22-26th 2016, Prague, Czech Republic. Available at: http://dx.doi.org/10.1145/2991041.2991043 (Accessed 3 October 2018.)

Metz, C. (2019) 'Businesses Will Not Be Able to Hide': Spy Satellites May Give Edge From Above. *The New York Times, 24 January 2019. Available at:* https://www.nytimes.com/2019/01/24/technology/satellites-artificial-intelligence.html (Accessed 6 February 2019).

Milne, D., & Watling, D. (2018) Big data and understanding change in the context of planning transport systems. *Journal of Transport Geography*. Article in Press. https://doi.org/10.1016/j.jtrangeo.2017.11.004

Murphy, K. P. (2013) *Machine learning : a probabilistic perspective*. Cambridge, Mass. [u.a.]: MIT Press.

Mäntysalo, R. & Kanninen, V. (2018). Kaupunkiseutujen integroiva suunnittelu ja hallinta. Yhdyskuntasuunnittelu 56: 3, 4–8.

Neoh JG, Chipulu M & Marshall A 2017. What encourages people to carpool? An evaluation of factors with meta-analysis. Transportation 44, 423–447.

Novak, J., Ahas, R., Aasa, A., Silm, S. (2013) Application of mobile phone location data in mapping of commuting patterns and functional regionalization: a pilot study of Estonia. *Journal of Maps,* Vol. 9(1)*:* 37–41. https://doi.org/10.1080/17445647.2012.762331

Oxford Internet Institute (2018) In Conversation with Viktor Mayer-Schönberger. Interview. YouTube. https://www.youtube.com/watch?v=5d6QabsQ2V0 (Accessed 27 September 2018.)

Ozga, J. (2009) Governing education through data in England: from regulation to self-evaluation. *Journal of Education Policy,* 24:2, 149-162.

Ozga, J. (2012) Governing knowledge: data, inspection and education policy in Europe. Globalisation, Societies and Education, 10:4, 430–455.

Parnami, A., et.al. (2018) Deep Learning Based Urban Analytics Platform: Applications to Traffic Flow Modeling and Prediction. In: ACM SIGKDD Workshop on Mining Urban Data (MUD3), London, UK. Available at: https://webpages.uncc.edu/mlee173/pdfs/mud3_18.pdf (Accessed 7 February 2019).

Pei, T., Sobolevsky, S., Ratti, C., Shaw, S.-L., Li, T. & Zhou, C. (2014) A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geography Information Science*, Vol. 28: 1988–2007. https://doi.org/10.1080/13658816.2014.913794

Puga, D. (2010) The magnitude and causes of agglomeration economies. Journal of Regional Science 50: 1, 203–2019.

Ren, Y. (2018) Big data clustering and its applications in regional science. In: Schintler, L. A., & Chen, Z. (eds.) *Big Data for Regional Science*. Oxon: Routledge: 257–264.

Rodríguez-Pose, A. (2018) The revenge of the places that don't matter (and what to do about it). *Cambridge Journal of Regions, Economy and Society*, Vol. 11: 189–209.

Santala, V. (2017) Avoimen datan kehitys pohjoisella kasvuvyöhykkeellä. Haasteet ja mahdollisuudet sekä merkitys kasvuvyöhykkeen kehittämiselle. (Translated: Development of open data in the North growth zone. Challenges and opportunities as well as the importance of developing the growth zone.) Master's Thesis. University of Turku Faculty of Mathematics and Natural Sciences, Department of Geography and Geology, 02-02-2017.

Sauv, D. (2018) *Cargo e-AWB Implementation Playbook*. IATA, September 2018. Available at: https://www.iata.org/whatwedo/cargo/e/eawb/Documents/e-awb-implementation-playbook.pdf (Accessed 6 February 2019.)

Schintler, L. A. (2018) The constantly shifting face of the digital divide: Implications for big data, urban informatics, and regional science. In: Schintler, L. A., & Chen, Z. (eds.) *Big Data for Regional Science*. Oxon: Routledge.

Sharma, Ray (2018) Telia Finland's Crowd Insights Helps Urban and Traffic Planning with Anonymized Mobile Network Data. *The Fast Mode*, 20 March 2018. Available at: https://www.thefastmode.com/technology-solutions/12295-telia-finlands-crowd-insights-helps-urban-and-traffic-planning-with-anonymized-mobile-network-data (Accessed 24.1.2019).

Shi, X.; Qian, Y. & Dong, C. (2017) Economic and Environmental Performance of Fashion Supply Chain: The Joint Effect of Power Structure and Sustainable Investment. *Sustainability* 2017(9): 961.

Shreenath, V. M. & Meijer, S. (2016) Spatial Big Data for Designing Large Scale Infrastructure: A Case-Study of Electrical Road Systems, *IEEE/ACM 3rd International Conference on Big Data Computing Applications and Technologies (BDCAT)*, Shanghai, 143-148. Available at: : https://ieeexplore.ieee.org/abstract/document/7877061 (Accessed 1 February 2019.)

Silm, S. & Ahas, R. (2010) The seasonal variability of population in Estonian municipalities. *Environmental Planning A*, Vol. 42: 2527–2546.

Silm, S. and Ahas, R. (2014a) The temporal variation of ethnic segregation in a city: evidence from a mobile phone use dataset. *Social Science Research*, Vol. 47(1): 30–43.

Silm S. & Tiru M. (2018) Transnational mobility between Estonia and Finland based on mobile phone roaming datasets. Presented at Annual meeting of Finnish geographers 2018.

Spiekermann & Wegener (2007) ESPON Project 1.4.4. Preparatory Study on Feasibility of Flows Analysis. Final Report.

Stead D & Meijers E (2009) Spatial Planning and Policy Integration: Concepts, Facilitators and Inhibitors. Planning Theory & Practice 10: 3, 317–332.

Stough, R. & McBride, D. (2014) Big Data and U.S. Public Policy. *Review of Policy Research*. Vol. 31 (4).

Tallon, P. P. (2013) Corporate Governance of Big Data: Perspectives on Value, Risk and Cost. *Computer,* Vol. 46. Iss. 6. June 2013.

Tang, J.; Wang, Y.; Wang, H.; Zhang, S. & Liu, F. (2014) Dynamic analysis of traffic time series at different temporal scales: A complex networks approach. Physica A: Statistical Mechanics and its Applications, Vol. 405: 303–315.

TA (2007) Territorial Agenda of the European Union. Towards a more competitive and sustainable Europe of diverse regions. Agreed on the occasion of the informal ministerial meeting on urban development and territorial cohesion in Leipzig on 24/25 May 2007.

TA (2011) Territorial Agenda of the European Union 2020. Towards an inclusive, smart and sustainable Europe of diverse regions agreed at the informal ministerial meeting of ministers responsible for spatial planning and territorial development on 19th May 2011 Go¨do¨ llo˝, Hungary.

Turck, M. (2018) Great Power, Great Responsibility: The 2018 Big Data & AI Landscape. 26 June 2018 *Matturck.com*. Available at: http://mattturck.com/bigdata2018/ (Accessed 20 September 2018)

Turck, M. & Obayomi, D. (2018) Big Data & AI Landscape 2018. Infographic. Available at: http://mattturck.com/wp-content/uploads/2018/07/ Matt_Turck_FirstMark_Big_Data_Landscape_2018_Final.png (Accessed 8 October 2018.)

Uyarra, E., Sörvik, J. and Midtkandal, I. (2014) Inter-regional collaboration in research and innovation strategies for smart specialisation (RIS3), S3 Working Paper Series No 06/2014. Available at: http://s3platform.jrc.ec.europa.eu/documents/20182/114903/JRC91963_Interreg_Collaboration_RIS3_final.pdf/89b7b310-8378-4e03-bcd9-2680061d037e (Accessed 3 October 2018.)

U.S. Dept. Transportation (2010) , traffic sensor costs table, https://www.itscosts.its.dot.gov/its/benecost.nsf/DisplayRUCByUnitCostElementUnadjusted?ReadForm&UnitCostElement=Inductive+Loop+Surveillance+at+Intersection+&Subsystem=Roadside+Detection+(RS-D) , accessed on May 07th 2019.

Van Asschea, K. & Djanibekovb, N. (2012) Spatial planning as policy integration: The need for an evolutionary perspective. Lessons from Uzbekistan, *Land Use Policy*, Vol. 29: 179–186.

Vanoutrive, T., Van De Vijver, E., Van Malderen, L., Jourquin, B., Thomas, I., Verhetsel, A. & Witlox, F. (2012) 'What determines carpooling to workplaces in Belgium: location, organisation, or promotion?' *Journal of Transport Geography*, Vol. 22, 77–86.

Vissim (2017) http://vision-traffic.ptvgroup.com/en-us/products/ptv-vissim/. Last accessed on May 8th 2019.

Wang, W.; Li, J.; Huang, F. & Feng, H. (2008) Design and implementation of log-Gabor filter in fingerprint image enhancement. Pattern Recognition Letters, 2008. pp. 301–308.

Wang, X.-H. & Xiao. J.-M. (2003) A radial basis function neural network approach to traffic flow forecasting. In: Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems.

Weiss, S.M. and N. Indurkhya (1998) *Predictive Data Mining: A Practical Guide*, Morgan Kaufmann Publishers.

Zijlstraa T & Vanoutrive T (2018) The employee mobility budget: Aligning sustainable transportation with human resource management? Transportation Research Part D 61, 383–396.

Zonneveld, W. & Trip, J.J. (eds.) (2003) Megacorridors in North West Europe. *Investigating a New Transnational Planning Concept*. Delft, Delft University Press.

Yuan, Y., M. Raubal, and Y. Liu. (2012) Correlating mobile phone usage and travel behaviour – A case study of Harbin, China. Computers, Environment and Urban Systems 36:118–30.

van Zuylen, J. H. (1978) "The information minimizing method: validity and applicability to transport planning", In: *New Developments in Modelling Travel Demand and Urban Systems* (edited by G. R. M. Jansen et al).

APPENDIX 1 – Interview Question Themes and Interviewees

Interviewees

One mid-level manager, Statistics Finland, 29.9.2018.

Three planning-related staff members, City of Helsinki, 27.9.2018.

A staff member from Regional Council of Uusimaa, 22.11.2018.

A staff member from Northern Growth Zone Initiative, City of Turku, 2018.

Interview Questions

Theme 1: Your work & organization, utilizing data, policy domains

1.1. Please tell us a little about your work. Briefly, what do you do for your job?

1.2. What is the purpose/objective of your organization in terms of data provision/utilisation?

1.3. What kinds of data are most relevant to your organization's activities?

1.4. What are some of the key ways your organization uses or produces data?

1.4a. Does your organization ever use data from other organizations? If yes, which organizations?

1.4b. Does your organization provide or share data with other organizations? Which ones?

1.5. Is your organisation utilising big data in its activities? If yes, how?

1.6. What are some key challenges your organization faces in using data/big data?

Theme 2: Corridor Development

2.1. What kind of data your organisation currently have/produces that somehow describe interactions of people, goods and ideas between actors in different administrative areas (statistical units)? e.g. interactions related to commuting/among companies, entrepreneurs, investors, and consumers etc.

2.2. Where do you see this development is going? What kind of data could be emerging/there already exists that would better capture these interactions and flows?

2.3. Does your own organisation play any role in providing this kind of data?

Theme 3: future possibilities and challenges in utilizing big data

3.1. What emerging sources of data, trends in data, data strategies, or data-weak signals do you believe will have a big impact on your organization? Why? What about developments related to big data?

3.2. What types of current problems do you believe you'd be able to address with some new kinds of data? What is that data?

3.3. From your perspective, what are some key emerging forms of data that will have a high-impact on your organizations internal policies and wider public policy-making? Why?

3.4. How is your organization building its capacities to analyze data and more consistently make decisions based on that data analysis throughout the organization?

APPENDIX 2

Table 9.1: Assessing Example Datasets by Applicability to Corridor Development

Dataset		**Reasons for applicability rating**	**Applic-ability Score**
Dataset Name	*To which interactions / flows is it relevant?*	*Qualitative assessment of the dataset's applicability to transnational corridor development in the study area of this targeted analysis.*	
Passive Mobile Positioning Data, Telia (see Case 3)	**Physical** flows of individuals **Social** patterns of mobility **Digital** communication flows among individuals	This dataset has high potential value to policymakers because of its frequent updates, fine-grained and deep temporal resolution, and usefulness in producing nearly up-to-date O-D matrixes for planners. Similar dataset exists for all nations in the research area and all nations in the EU, meaning it could be used for cross-border analysis. However, its coverage is limited because it is only one company's data out of a handful operating in Estonia. This limit can be overcome by obtaining and combining data from more operators. Where the data is available, it is usually for a price. Telia also offers flow analytics services to public sector. A benefit to this data is that analytical methods for producing policy relevant insights are highly developed based on many years of research. Therefore, it is easier for other nations develop everyday mobility databases and to harmonise them to span corridors.	High
Traffic intensity data detected by induction loop sensors (see Case 1)	**Physical** flows of vehicles	Data of this kind is produced for many of the major highways across Europe. It is a good candidate for studying functionality of Europe-wide corridors, e.g., the TEN-T network. However, its exact form and ways of being accessed vary by nation in our study area and presumably throughout Europe, so a harmonisation of how it is presented for use is needed. Because the sensors it relies upon are largely reliable, the datasets they produce are comprehensive and can be deep, going back many years, with a fine-grained temporal resolution. The data can also be made accessible in real-time, which introduces opportunities for forecasting. Despite these positive attributes, case 1 of this targeted analysis found that pre-processing this kind of data is time intensive and the value of the O-D Matrixes that can be produced from it are not yet clear for policy-making, thus requiring further investigation. However, if this type of data were combined with other forms of data, it could serve as a base-level of data upon which other temporally varying data could be explored and mapped. Predictive models are among the toolset that policymakers can utilize with researchers to leverage the data for many applications. However, a variety of frequently updated economic indicators are needed before the full power of traffic models can be harnessed.	Moderate
Project and beneficiary data from EU Interreg and Interact HIT (see Case 2)	**Social** interactions of project collaborators	This data covers the whole EU and includes data for more than half of the projects in the latest reporting period. It is publicly available on Keep.eu but requires some effort to download and prepare it for use. The visualisations that can be produced, however, provide clear insights as to how the funding instruments function to connect actors in different regions, and illustrates how growth corridors may be involved in collaboration partnerships. While the insights produced hold	Moderate

		some value for policymakers who design Interreg and Interact funding instruments, the wider value for all organisations involved in corridor development is less obvious because the dataset is too narrow. However, if it were combined with datasets for additional funding instruments, a richer picture of spatial links created through project collaborations could be produced.	
Instagram Posts	**Digital** interactions with physical places or objects	As a popular social media platform, geotagged data exists for photos taken across Europe. Instagram posts can only ever serve as a proxy data, as not all people are active users, and the demographics of the user base skews toward specific generation. However, it can provide insights regarding what locations are considered photogenic, what languages are most used around key touristic sites, or what activities are most popular in certain places. Because Instagram is a company, it can change its policies regarding accessing its data via its API at any time. Instagrem can only provide a limited level comprehensiveness regarding a topic of inquiry.	Moderate
All digital artifacts from a planning process (e.g. Eräranta 2019)	**Social** interactions among individuals involved in a planning process	A digital artefact can be an architectural rendering, meeting minutes, feedback compiled on a design, etc. that is kept by planners in a planning process. This data is only available on a case-by-case basis and its quality varies from locality to locality. It can shed light when and in what parts actors involved in a planning process are engaged that process. Revealing such patterns can help in changing the ways of working and possibly policies regarding participation in planning, but at this point, while quite interesting, it provides a limited value for the effort required to use it.	Low
Company Board Members Data (e.g. Kalliomäki et al. 2018)	**Social** interactions between company board members	This kind of data exists for all nations in the study area and probably in all EU nations. The data sometimes is provided for a cost. Some extraction effort is required to make it useful and proper analysis of it requires partnering with researchers. For policy-makers, analysing the overlap and differences among company board members in relation to temporal and spatial data can provide useful insights regarding the geographical distribution of decision-making and strategic connectedness of areas through company boards. This kind of knowledge can be utilised e.g. in designing further networking activities. However, further examination is needed how these types of insights can be interpreted into policy for corridor governance.	Moderate
Copernicus Satellite Data	**Physical** land use changes.[29]	Copernicus program aims to serve the needs of policymakers across Europe and has a well-established sensor (both space-based and terrestrial) and data ecosystem. The Copernicus Land Monitoring Service can provide policymakers with up-to-date information about physical and spatial changes, e.g. the growth of built-up land area over time, making it an ideal data source for transnational corridors. These types of insights can help policymakers guide the production of new built-up spaces. The data is available for free but requires skills to use it effectively. Therefore, public entities would likely need to partner with researchers or vendors to make use of this dataset.	High

[29] https://land.copernicus.eu/user-corner/land-use-cases (Accessed 8 June 2019)

APPENDIX 3 – Illustrative List of Tools for Data Analytics

Table 0.2: Examples of AI-based data analytic services[30].

Tool Name	What it Does	Learning Curve
Rapid Miner https://rapidminer.com/	Supports the whole cycle of prediction modelling from preparing data to deployment.	The interface is based on block diagram approach with blocks pre-programmed with certain functions. The user interface is quite easy to understand and the provided tutorials make it fast to get started. However, to grasp how to fully use the tool and interpret the results, extensive training and learning is required.
BigML https://bigml.com/	Individual or enterprise-level users can compose complex AI tools built from basic AI components.	This platform offers a unified environment for conducting data analytic research. Its training videos are comprehensive and cover the necessary information to use the tool effectively. Several product packages target individual users and enterprise users.
Google Cloud AutoML https://cloud.google.com/automl/	Allows users to "Train high-quality custom machine learning models with minimum effort and machine learning expertise."	Users can access AutoML Translation, Natural Language, and Vision tools. Vision, for example makes it easy to build models based on Google's image recognition neural networks. Users can upload pictures to further train these networks to specific needs.
Paxata https://www.paxata.com/	Paxata is a data science collaboration tool targeting the enterprise users. It focuses on data governance, collaboration, and distributing data insights throughout the organization.	Because of it targets enterprises, Paxata requires a set up process. The company claims it as easy to use for end-users in an organization as a spreadsheet while sophisticated enough for skilled data science teams.
Trifacta https://www.trifacta.com/	"Wrangling" data, Trifacta aims to make preparing raw data for analysis easier and more standardised for organizations.	Trifacta targets enterprise clients, but offers packages for individuals and groups. The userbase it targets are analytics executives, IT leaders, and data engineers and analysts.
Microsoft Azure Machine Learning https://azure.microsoft.com/en-us/services/machine-learning-service/	Marketed as an end-to-end solution. Data can be cleansed and prepared using Azure ML and then data can be split into training and test sets. Different algorithms can be used to train the model. After this the models can be deployed at cloud scale.	Azure ML Studio is a browser-based machine learning platform with a drag-and-drop interface. Without coding, users can leverage powerful Machine Learning functions. Comprehensive tutorials and sample experiments are available to help new users learn the system.

[30] Examples of AI-based data analytic services that require no or little experience to use. The illustrative list was compiled in July and August 2018 to see what kinds of tools are becoming accessible to individuals and organisations.

ML Jar https://mljar.com/	ML Jar is a browser-based platform for quickly building and deploying machine learning models. It has an intuitive interface and allows you to train models in parallel.	A user can build a data model in three steps: 1) Upload your dataset; 2) Train and tune many Machine Learning algorithms and select the best one; and 3) Use the selected model to generate predictions and share your results.
Amazon Lex https://aws.amazon.com/lex/	Simplifies the process of building an audio chatbot for Amazon Alexa.	The user interface is highly intuitive making it easy to program a chat bot which can access other AWS functions.

ESPON 2020 – More information

ESPON EGTC
4 rue Erasme, L-1468 Luxembourg - Grand Duchy of Luxembourg
Phone: +352 20 600 280
Email: info@espon.eu
www.espon.eu, Twitter, LinkedIn, YouTube

The ESPON EGTC is the Single Beneficiary of the ESPON 2020 Cooperation Programme. The Single Operation within the programme is implemented by the ESPON EGTC and co-financed by the European Regional Development Fund, the EU Member States and the Partner States, Iceland, Liechtenstein, Norway and Switzerland.