Challenges of Combining Open and Commercial Data Sources in Visitor Mobility Estimations

J. Grönman*, P. Rantanen*, P. Sillberg*, T. Pohjola†, T. Jönkkäri*
* Tampere University/Faculty of Information Technology and Communication Sciences, Pori, Finland
† Turku School of Economics at the University of Turku, Pori Unit, Pori, Finland
petri.rantanen@tuni.fi

***Abstract*—This paper explores the opportunities and challenges of open data. Open and free interfaces can provide useful data, but sometimes there are challenges with data sufficiency and quality. Open data can be enriched by acquiring commercial data, and by merging multiple data sources, a more comprehensive understanding of a situation can be obtained. The focus of this paper is on nature parks, but the technical solutions can be applied to other venues. In the case presented by this paper, the parks are equipped with on-location visitor counters, and with the help of commercial data provided by a mobile network operator, and open sources, such as the weather and road traffic services, we are attempting to gain a better view of visitor movements. The motivation for this research originated from the cities located in the Satakunta region of Finland. For the purpose of developing recreational and business opportunities in the region, the cities require a better understanding of movement of people - which locations and services are popular, and which locations need improvement. To help achieve this goal, this paper presents a technical solution of how analysis and visualization of combined data can be used to estimate visitor amounts within a geographical area.**

***Keywords*—open data, commercial data, visualization, mobile network operators**

I. INTRODUCTION

This paper presents the measures taken in the SmartMove project, which wanted to promote with concrete solutions the tourism and service companies of the Satakunta area benefiting from the rapidly growing local tourist flows due Covid-19 pandemic in cultural and leisure destinations and routes. By improving intelligent visitor information management and communication, it would be possible to predict and promote the healthy and safe use of services in almost real time, avoiding traffic jams and unnecessary human contact. At the same time, it increases the sense of security to move outside in cultural and leisure destinations. The social isolation associated with Covid-19 pandemic has significantly increased people's desire to spend time outdoors and in nearby destinations, including hiking and moving around in nature and cultural destinations [1].

The popularity of outdoor activities has been reported in many countries. According to [2], "there has been so much traffic in the most popular camping destinations that the roadsides have been blocked with parking and the police have been needed to investigate the matter. The shelters have also been crowded", which can feel particularly frustrating in exceptional circumstances and reduce the feeling of safety.

The tourism profiles of both Finland and Satakunta rely heavily on nature and outdoor environments. Satakunta's natural and cultural destinations have significant potential in the new world situation to develop safe and attractive digital solutions to promote smart and sustainable tourism destinations.

Smart solutions and various data sources will offer completely new development opportunities for travel destinations and attractions, as well as for businesses. For example, from a hiker's point of view, the problem is that when planning your own route, there is no information whether there are already many people on a certain route, rest stop or shelter, or whether the locations are free of other users. Real-time service data, or even indicative open statistical data on the number of users, visit times or congestion peaks of various outdoor attractions are not available in the Satakunta area or nationwide in Finland. Automatic information about non-congested, low-congestion or very congested destinations gives not only the user, but also tourism and leisure developers real-time information on, for example, how to schedule service capacity, recommend routes and carry out distribution in order to minimize congestion.

Nature destinations and rural areas play a crucial role in promoting the well-being of citizens and providing recreational opportunities [3], [4] such as hiking, biking, and admiring scenic landscapes also for the visitors. The advantages of nature destinations in enhancing physical and mental wellness are becoming more evident [5], [6], with studies showing how exposure to diverse ecosystems positively affects our microbiota and immune system [7]. In Finland, nature tourism is centred around forests, marshland, Lapland fells, lakes, and the Baltic Sea. Finland's "Everyman's rights" allow individuals to walk, ski, cycle, and camp in all forests and engage in activities such as swimming, boating, and fishing in the waters. During the winter, individuals can walk, ski, and skate on ice.

Access to place-based rich data and smart services can further increase the appreciation of nature-based destinations among new audiences and deepen the exposure to various cultural experiences. This information and immersive contents can lead to the discovery of new locations for relaxation, recreation, education, immune system boost,

and more. In addition to nature itself forest areas possess cultural and industrial heritage that appeals to tourists and locals alike [8]. Service providers can use data to strengthen the bond to these places [9] and make the experiences even more meaningful [10], [11]. Nature and rural areas, as well as more urban environments, can be considered as hybrid spaces in which tourists and other stakeholders create value together in the development of smart tourism [12]–[14] by blending physical space with real-time interactions [15]–[17]. The use of big data sheds light to visitor behaviour and additionally IoT sensor data can further improve the long-term sustainability of nature and rural tourism operations and stakeholder networks [11], [18], [19] by managing tourism flows in sensitive areas [12].

In this paper, a use case is presented, which combines open data sources and commercial mobile operator data for calculating and estimating visitor amounts in the nature parks of the Satakunta region of Finland. In the following sections, the available data - both from open and commercial sources - is described and illustrations on data visualization are presented.

Therefore, in an attempt to unravel the mysteries of smart tourism, the rest of this paper has been divided into the following three sections: II describes the use case selected for this paper, the data sources utilized, and methods for visualizing the data; III explores the challenges faced and possibilities discovered during the research; and finally, IV gives a summary of the main findings.

II. USE CASE: ESTIMATING CONSUMER MOBILITY IN SATAKUNTA REGION

The use case described in this section presents a map-based visualization tool for smart destination development. In essence, the goal is to collect historical data from multiple sources and combine the data to produce an overview of past visitor movements in an area, and estimates for possible future visitations. The primary users of the tool are the entities responsible for the development of the Satakunta region (cities, municipalities) and the companies and organizations responsible for managing nature parks. The tool could also be utilized by users, who are interested in discovering the various nature locations in the region, or companies interested in establishing new businesses. The focus is on nature parks, but the technical solutions could be applied to other venues as well.

A. High-level Service Architecture

The overall technical view of the system and its connections are presented in Fig. 1. The tree main parts are the two clouds (representing Internet) and a folder representing the containerized `SmartMove Service`. Open source implementation (i.e., source codes) of the service is available at [20].

Internet (i.e., clouds) allows the service to retrieve data from 3rd party APIs (such as Wikipedia). It also provides

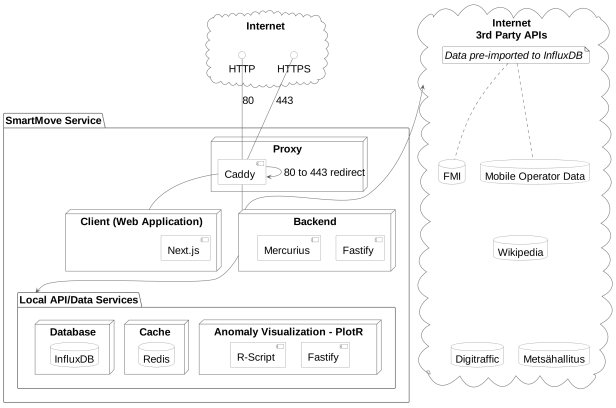

Fig. 1: High-level diagram of the service architecture.

the entry-point for the users of the service to access the system, mainly over HTTP/S connections.

Inside the service package, each node (represented by rectangular cuboids) is a Docker container. The main nodes are `Proxy`, `Client`, and `Backend`. The supporting nodes (Database, Cache, and Anomaly Visualization) are separately packaged in `Local API/Data Services`. It means that any of those are designed to be run independently if needed to.

The first node, `Proxy`, is implemented by *Caddy* server. In essence, it simply delivers the requests appropriately between `Client` and `Backend`. The second node is `Client` that provides the visual user interface for the consumers of the service. It is implemented using Node.js runtime and *Next.js* framework. The `Backend` also utilizes Node.js runtime. It is built on *Fastify*, a low overhead web framework, and *Mercurius*, which is a GraphQL adapter for *Fastify*. Both the `Client` and `Backend` communicate with each other and 3rd party services by using the GraphQL query language. Additionally, the *Mercurius* component in `Backend` enables caching by utilizing *Redis* in-memory cache, making subsequent unmodified queries near instantaneous.

The service does not have a dedicated database as it just shows the information in read-only mode. It does, however, utilize databases such as InfluxDB through other supporting services. For example, we implemented a tool for anomaly visualization as a micro service, `PlotR`. It is useful for inspecting temporal data on side-by-side plotted diagrams. `PlotR` is described more specifically in section II-D.

B. Utilization of Open Data

The main user interface of the service can be seen in Fig. 2. The example shows all data contents selected with the exception of mobile network operator data, which is illustrated separately in Fig. 4, and explained in more detail in Subsection II-C. By clicking the sliders, the users can choose which point of interest (POI) - i.e. the data sources - are visible on the map, and by clicking the POIs, the users can see the data of the clicked POI. The

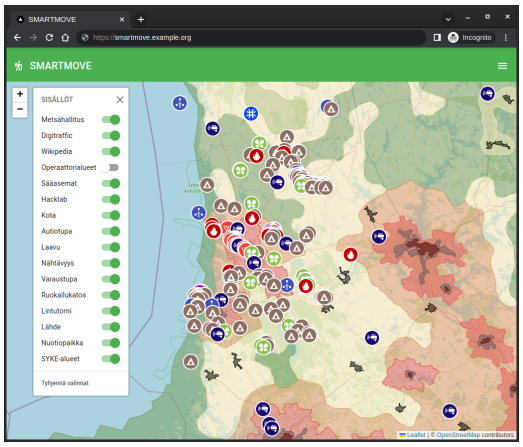

Fig. 2: Example visualization of open data sources.

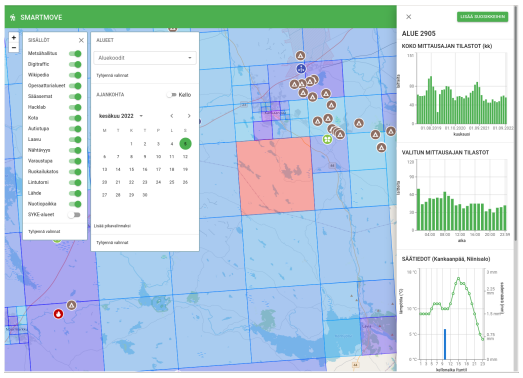

Fig. 3: Example visualization of mobile network operator data with weather and visitor statistics.

majority of the POIs provide general information of the area (places to see, the Wikipedia article of the location, etc.). In the context of this paper, the most interesting ones are Digitraffic and Sääasemat (weather stations of the Finnish Meteorological Institute, FMI [21]). Digitraffic [22] is a service managed by the Finnish Transport Infrastructure Agency, which provides open data on the road, rail and waterway systems of Finland. In this case, Digitraffic's APIs could be used, for example, to retrieve traffic volumes (the amount of cars) passing on a particular highway.

By clicking the map, the user can also see data of a specific area (grid block), as illustrated in Fig. 3. In the example figure the selected area is highlighted in red, and the desired data can be seen in a scrollable list on the right side of the user interface. In this case, the data shows (from top to bottom): the monthly visitor amounts for the whole measurement period; the hourly visitor amounts for the date (June 4th, 2022) selected from the calendar widget; and the weather history for the selected date. The visitor amounts are calculated based on the mobile operator data (described in more detail in the following subsection II-C, and the weather data is retrieved from FMI's open data service. Thus, the user can, with a quick glance, see an overview of visitors on a particular time in relation to the weather.

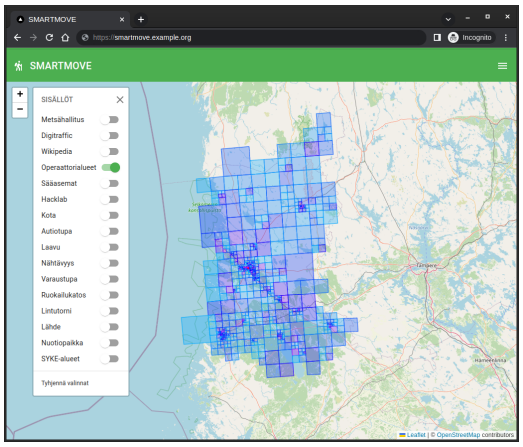

Fig. 4: Example visualization of mobile network operator data.

Of course, especially weather can be a complicated relation to figure out by simply looking at the figures as "poor weather" is relative to the location and to the time of the year (season). Cold weather in the winter near to a ski center can be desirable, but not near a beach in the summer. Similarly, in the case of time, certain locations receive higher visitor amounts during holiday seasons. Unfortunately, at least in Finland, there does not seem to be a programmably accessible API, which could be used to retrieve public holidays and holiday seasons for a particular year, requiring either hard coding the dates or scraping the information from web pages - both being somewhat unreliable or high maintenance solutions in the long run.

C. Utilization of Commercial Data – Operator Data

Today, almost everyone has a cell phone. Thus, at least in principle, the information about the location of the phones, could be a great asset in improving visitor counting accuracy. In Finland, a couple of operators offer commercial services, which provide anonymized data about mobile devices. In general, the service consists of a dataset, which contains per area device amounts with accuracy that vary between mobile network operators. The accuracy can differ both in time and area sizes. In our dataset, the device amounts were reported hourly, but for other operators, the amounts could be reported on a daily basis (once every 24 hours). The dataset contains the amounts for devices that have subscription for the operator in question, and an estimate of other operators' devices in the area. The total amounts are calculated based on a sample collected by the operator, and then extrapolated based on the amount of population living in the area in question. Foreign subscriptions (e.g. SIM cards) are excluded from the dataset. Regardless, at least in principle, ordering the service from one operator should be enough for getting a fairly good estimate on the device amounts in the area, though we did not perform any multiple operator analysis to find out how accurate the promised estimations were.

An example grid of Satakunta region, as provided for our dataset, can be seen in Fig. 4. The region - which has the combined land and sea area of approximately the size of the countries of Jamaica or Kosovo - is divided into 670 smaller areas (blocks), with sizes between 500x500 meters and 20x20 kilometers. In the grid visualization, blocks with lighter (blue) color represent lower device amounts and darker (blue) color represent higher device amounts on the selected time period (in the example, the entire data gathering period of three years). It is important to note that the blocks do not necessarily match a single specific cell tower, but are aggregate products of any number of cell towers, with the actual number not provided in the dataset for privacy and security reasons. The size of a block also depends on the operator's estimate of total device amount in the area. The higher the amount of devices, the smaller (more accurate) the block is defined to be. Thus, it can be assumed that the larger blocks have a lower amount of devices, in general. Furthermore, in our dataset, any geographical area has less than five (5) devices, the data has been omitted for privacy reason by the mobile network operator, though in our use case, the region has sufficiently high population density to generate data on every block. The Finnish mobile network operators also provide data for other countries they operate in (e.g. Estonia, Sweden and Norway, depending on the operator), and these countries have their own (legal) regulations, which may affect the available data accuracy and quality. It is also possible to get estimates for the device movements (e.g. how many devices per day moved from one block to another), but this information turned out to be too coarse for reliable estimation of visitor amounts.

D. Visualization Methods

There was a growing interest toward having more visualizations and automatically computed analysis of the available data. The chosen specimens for this experiment were the weather data and the mobile operator data. Being able to do experiments and seeing (i.e., visualizing) how the data behaved was a major driving factor. The second step was to apply k-means clustering to find out whether there were any automatically discoverable classifications (i.e., cluster centroids) in the data. Unfortunately, due to limited resources only the first part, visualization, was successfully implemented in the user interface. However, the k-means clustering could be implemented and executed as one-shot offline analysis of the complete data set.

As was mentioned above in Fig. 1, the PlotR is implemented as a micro service. It uses Node.js runtime together with Fastify for providing a web service. The web service is a simple facade to pass through the URL query parameters to R-Script which makes computations that generate a downloadable diagram image.

Fig. 5 shows a zoomed screen capture of the bottom part of the scrollable list which was initially seen in Fig. 4. The user interface in Fig. 5 has been manually translated to English for the convenience of the reader.

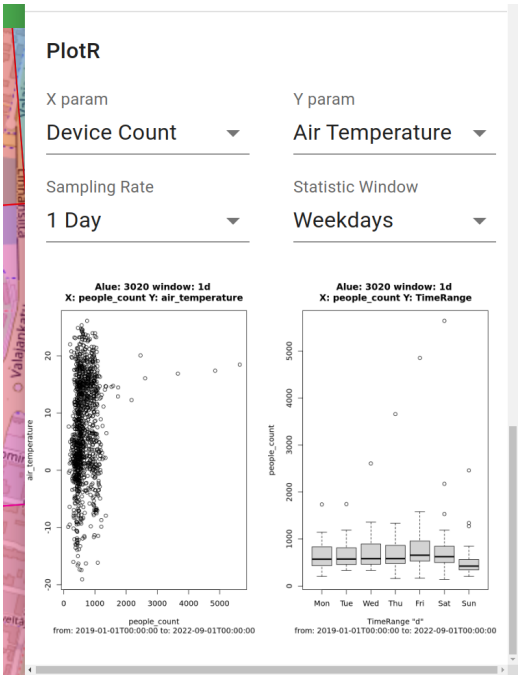

Fig. 5: Anomaly visualization with PlotR micro service.

The user interface consists of two parts: 1) the variable configuration, and 2) the diagram image, which is further divided to the XY scatter plot on left, and the Xt box plot on right.

Each grid block can be selected to display the corresponding detailed plot visualization. It is possible for the user to customize each plot using four additional parameters: `X param`, `Y param`, `Sampling Rate`, and `Statistic Window`. The default settings for these are chosen to be `Device Count`, `Air Temperature`, `1 Day`, and `Weekdays`, respectively.

The sample rate effects both diagrams by dictating the amount of data points in the dataset. It has seven options ranging from one hour (native sampling in this case) up to one year (aggregated samples). Parameters for X and Y axles include 13 similar options. Axis X affects both diagrams while the Y axis affects only the scatter plot on the left. The statistic window dictates how the dataset will be grouped on the box plot on the right. It has six options for grouping up the data; from hourly to yearly groups.

According to the shown example diagram in Fig. 5 it seems to indicate a few features: 1) the bulk of the device count is averaged to approximately 600 devices (both plots); 2) each weekday except for Sundays are quite similar (box plot); 3) there are a few outlier data points that show very high device count over several days, perhaps there was a popular public event (both plots); 4) higher air temperature might indicate slightly higher device count (scatter plot). However, one figure only can not explain everything with certainty, thus further experimentation with the data is needed to reveal more insight into the observed data.

III. DISCUSSION

The primary motivation for considering the utilization of mobile network operator data was the known limitations of the available open data sources. Open data, in general, is provided as general datasets, and it is often the task of the user (or developer) to figure out the potential use case scenarios. This also means, that the data may not be quite what is required. In our use case, the data was not actually measured at the nature parks. The historical weather data gathering stations around the Satakunta region are relatively close to the nature parks, but this may not be the case in larger and more rural nature areas. Similarly, the traffic data - at least in Finland - is collected mainly on the main highways. Thus, if the highway does not directly go to the nature park, but, for example, travels between two major cities with only a smaller side road to the park (or multiple intersecting roads to other destinations), it becomes exceedingly difficult to figure out the actual destinations of the travellers.

Metsähallitus (the Finnish Forest Administration), which is an agency that amongst other duties manages the nature parks in Finland, also provided us with visitor statistics for the parks, but after discussions with the representatives of the agency, it was discovered that the data are mostly estimations, and not actual measured on-location. In Finland, as well as in many other countries, nature parks are free and open for everyone, which is a good thing for visitors, but bad for counting exact visitor counts as there are no ticket sales to count. It is often possible to enter and exit the parks in multiple locations and there are several nature trails one can follow, and only a limited amount of these routes have physical visitor counter devices. The challenges of providing electricity and (winter) weather proof devices to nature areas often technically limit the counters to simple infrared counters, as opposed to, for example, camera-based methods. Also, it is worth noting that even though the use of the visitor statistics API was provided for us free of charge, the API is not as such open to everyone, and the access must be negotiated with Metsähallitus. With other venues, especially with those operated by companies, it is quite common, that visitor (or client) amounts are not openly available.

Mobile operator data is available in all locations, that have cell phone coverage. In some countries large nature areas may have coverage issues, but in Finland, and in Europe in general, there is coverage even in remote areas. As we are only interested in counting the devices that have at some point connected to the cell towers, technically even very poor network coverage is acceptable. Thus, the limiting factors come from the grid and block sizes defined by the operator (accuracy), from the legal regulations of the target area or country, and finally, from the cost of the data. In Finland, the data services provided by the operators are primarily aimed at companies, and to the best of our knowledge the data cannot be acquired freely or openly. Thus, the overall pricing options are reasonable for larger companies, but the cost might be a bit steep for a one-person startup or for someone who simply wants to try-out what could be done with the data. Especially, if the use case requires something special to negotiate, such as customizations to the provided grid or data accuracy.

For our use case of nature parks, the accuracy of the mobile device data was sufficient, but in high population density areas the smallest block size (500x500 meters) might be too large for some use cases. One challenging case could be to pinpoint specific shops in city centers. The grid also has a relatively high number of blocks, and in practice it might be necessary to focus on a smaller geographical area. It is also crucial to map any factors in the target area, which could affect the device amounts, often even if the data does not show any specific irregularities. For example, proximity of a large sports stadium, music festivals, or an arrival of a large cruise ship can temporarily cause huge increase in device amounts. The last two events are especially interesting as they bring in new people, and from the region's point-of-view it would be important to find out whether this "new people" simply stay "at the venue" or do they visit other businesses as well? Unfortunately, discovering these factors is not always trivial. In the Satakunta region, near the Jämijärvi nature park, there is a large military base, which (for obvious reasons) does not publish information about their daily activities. Nevertheless, these activities can affect the device data collected around the nature park's area.

Finally, as our mobile network operator data was collected between January 2019 and August 2022, it shows per-year (and per-season when comparing seasons of individual years with each other) variations in many locations, presumably caused by the COVID-19 pandemic travel restrictions. The parks are by nature large outdoor areas, and are thus somewhat excluded from the most serious effects of the pandemic restrictions. In any case, it would require a much longer data period, covering the pre-pandemic era, to estimate the effect of the pandemic on the visitor amounts. From a technical point-of-view, the data does not cause problems for analysing the visitor data during the pandemic, but it makes estimating future post-pandemic visitor amounts more-or-less impossible.

Furthermore, even though technically the device data analysis can be expanded to venues outside the nature park context, and to countries and areas outside of Finland, a large scale utilization would require further studies on the availability (and cost) of mobile network operator data on other counties. As a summary, the mobile operator data shows promise, but the applicability depends on the (commercial) data services available in the desired geographical target area.

IV. CONCLUSION

This paper presented a use case, in which a combination of open data sources and commercial mobile operator data was utilized to study visitor amounts in the nature parks of the Satakunta region of Finland. The available

data was described and the paper illustrated how the gathered datasets could be visualized. The existing open data sources were found to be somewhat limiting in terms of accuracy and availability. The mobile operator (device) data was sufficient for our use case, but certain limitations of the approach were also discussed in this paper. Overall, the mobile operator data showed promise, but the applicability depends on the (commercial) data services available in the desired geographical target area and venue.

ACKNOWLEDGMENT

This study has been funded by National funding instrument (AKKE) of Regional Council of Satakunta. The mobile operator data was provided by Telia Finland Oyj.

REFERENCES

[1] C. Alba, B. Pan, J. Yin, W.L. Rice, P. Mitra and M.S. Lin, "Covid-19's impact on visitation behavior to us national parks from communities of color: evidence from mobile phone data," *Scientific Reports*, vol. 12, 2022.

[2] Yle - The Finnish Broadcasting Company, "Korona voi levitä myös retkipaikoilla - Metsähallitus varoittaa ruuhkista ja ruohikkopaloista (in Finnish) [Corona can also spread at camp sites - The Finnish Forest Administration warns of congestion and grass fires (title translated into English)]," https://yle.fi/aihe/artikkeli/2020/03/26/korona-voi-levita-myos-retkipaikoilla-metsahallitus-varoittaa-ruuhkista-ja, 2020, accessed: 2023-01-17.

[3] S. Bell, L. Tyrväinen, T. Sievänen, U. Pröbstl and M. Simpson, "Outdoor recreation and nature tourism: A european perspective," *Living Reviews in Landscape Research*, vol. 1, no. 2, 2001.

[4] R. Puhakka, K. Pitkänen and P. Siikamäki, "On certain integrals of lipschitz-hankel type involving products of bessel functions," *The health and well-being impacts of protected areas in Finland*, vol. 25, no. 12, pp. 1830–1847, 2017.

[5] Y. Joye, R. Pals, L. Steg and B. Lewis-Evans, "Correction: New methods for assessing the fascinating nature of nature experiences," *PLoS ONE*, vol. 9, no. 1, 2014.

[6] T. Hartig, M. Mang and G.W. Evans, "Restorative effects of natural environment experiences," *Environment and Behavior*, vol. 23, no. 1, 1991.

[7] L. Ruokolainen, J. Lehtimäki, A. Karkman, T. Haahtela, L. von Hertzen and N. Fyhrquist, "Holistic view on health: two protective layers of biodiversity," *Annales Zoologici Fennici*, vol. 54, pp. 39–49, 2017.

[8] K. Kirillova, X. Lehto and L. Cai, "What triggers transformative tourism experiences?" *Tourism Recreation Research*, vol. 42, no. 4, pp. 498–511, 2017.

[9] L. Birnbaum, C. Wilhelm, T. Chilla and S. Kröner, "Place attachment and digitalisation in rural regions," *Journal of Rural Studies*, vol. 87, pp. 189–198, 2021.

[10] D. Buhalis and Y. Sinatra, "Real-time co-creation and nowness service: lessons from tourism and hospitality," *Journal of Travel and Tourism Marketing*, vol. 36, no. 5, pp. 563–582, 2019.

[11] T. Pohjola, A. Lemmetyinen, and D. Dimitrovski, *Value Co-creation in Dynamic Networks and E-Tourism*. Cham: Springer International Publishing, 2020, pp. 1–23. [Online]. Available: https://doi.org/10.1007/978-3-030-05324-6_92-1

[12] U. Gretzel, C. Koo, M. Sigala and Z. Xiang, "Smart tourism: Convergence of technologies, experiences and theories," *Electronic Markets*, vol. 25, pp. 175–177, 2015.

[13] D. Buhalis and A. Amaranggana, "Smart tourism destinations," in *Information and Communication Technologies in Tourism 2014*, Z. Xiang and I. Tussyadiah, Eds. Cham: Springer International Publishing, 2013, pp. 553–564.

[14] A. Caragliu, C. Del Bo and P. Nijkamp, "Smart cities in europe," *Journal of Urban Technology*, vol. 18, no. 2, pp. 65–82, 2011.

[15] D. Buhalis, "Technology in tourism - from information communication technologies to etourism and smart tourism towards ambient intelligence tourism: a perspective article," *Tourism Review*, vol. 75, no. 1, pp. 267–272, 2019.

[16] U. Gretzel, L. Zhong, C. Koo, A. Morrison and A. Morrison, "Application of smart tourism to cities," *International Journal of Tourism Cities*, vol. 2, no. 2, pp. 216–233, 2016.

[17] M.V. Ciasullo, O. Troisi and S. Cosimato, "How digital platforms can trigger cultural value co-creation? – a proposed model," *Journal of Service Science and Management*, vol. 11, pp. 161–181, 2018.

[18] C. Koo, L. Mendes-Filho and D. Buhalis, "Guest editorial," *Tourism Review*, vol. 74, no. 1, pp. 1–4, 2019.

[19] K. Boes, D. Buhalis and A. Inversini, "Smart tourism destinations: ecosystems for tourism destination competitiveness," *Tourism Destinations: Ecosystems for tourism destination competitiveness*, vol. 2, no. 2, pp. 108–124, 2016.

[20] DigiluontoSatakunta, "Smartmove OSS," https://github.com/DigiluontoSatakunta/smartmove-oss, 2023, accessed: 2023-01-17.

[21] Finnish Meteorological Institute, "The Finnish Meteorological Institute's open data," https://en.ilmatieteenlaitos.fi/open-data, 2023, accessed: 2023-01-17.

[22] Finnish Transport Infrastructure Agency, "Digitraffic," https://www.digitraffic.fi/en/, 2023, accessed: 2023-01-17.