Urban Planning for Active and Healthy Public Spaces with User-Generated Big Data

Loes van Renswouw[1], Sander Bogers[2], Steven Vos[1,2,3]

1 Fontys University of Applied Sciences, School of Sport Studies, The Netherlands
2. Eindhoven University of Technology, The Netherlands
3. KU Leuven, Policy in Sports and Physical Activity Research Group
loes.van.renswouw@fontys.nl, s.j.a.bogers@tue.nl, steven.vos@fontys.nl

Abstract

This paper explores the value of user-generated big data for urban planning of active and healthy public spaces. The research is situated in and focuses on Eindhoven, an innovative and sports-minded city in the Netherlands. Based on running data collected by two popular apps in The Netherlands and Belgium, we present three iterations that set out to gain understanding in what factors define good running environments. The first iteration uses data visualisation techniques to get geographic insight in our data, to identify running hotspots and other points of interest for further analysis. The second iteration uses a mixed method approach to combine running data with qualitatively scored environmental characteristics of the selected locations from iteration one to identify possible influencers of the attraction of these areas for runners. As it became clear that this approach requested further scaling, in the third iteration we explore how we can come to factors that are worth scoring. Creating a larger set of locations with a reduced number of variables allowed for more substantial statistical analysis. This approach helped to provide an initial insight in the relevance of some of the environmental factors for optimised running climates.

Introduction

Our society is facing a noticeable increase in sedentary behaviour and physical inactivity, which is a growing public health concern (e.g. Blair, 2009; Vos, 2016; World Health Organization, 2010). Policy making for public environments to promote healthy and active lifestyles is a contemporary topic in both government practice as across multiple disciplines in research (e.g. interaction design, urban design, geography, sociology and psychology (Frumkin, 2003)). There is ample evidence that individual, social and environmental factor influence physical activity levels and behaviours (e.g. Cleland et al., 2012; Granner et al., 2007; Larson et al., 2014; Mansfield et al., 2012; Sallis et al., 2012). Furthering one of the grand societal challenges of increasing physical inactive lifestyles in most of the western world, the research fields see this trend being influenced by how the space around us is shaped (World Health Organization, 2010). In this paper the focus will be the role of the physical environment in promoting physical activity, as research has shown that characteristics of built environments are related to rates of chronic disease. More specifically, physical activity is perceived as a critical mechanism in this (Sallis et al., 2012). Because physical activity in urban areas often happens in public space, local and national governments can provide focus on the health values through their urban planning. Although larger cities and metropoles have not necessarily developed keeping these values in mind, health oriented environments stimulate the wellbeing, happiness and welfare of the people using it (Sallis et al, 2016).

The purpose of this paper is to explore the possibilities of user-generated big data for the urban planning of active and healthy environments, with the ambition of providing recommendations for physical activity policy making in the future. We explore these possibilities by means of a case study of recreational running in urban areas. Today, running is one of the most popular sport activities with 50 million participants in the EU-28 (Breedveld et al., 2015). Moreover, running is an expression of the exponential growth of unorganised sport practices in urban areas and parks, which require no or limited specific infrastructure (Vos et al., 2016; Borgers et al., 2016).

Where, how and why people run is notably influenced by urban layouts. In this paper we aim to create an insight in the extents of this influence, using GPS trail data collected via a popular running app among novice runners. We will discuss how this type of user-generated data can or cannot contribute to defining what makes a good running environment and by extension what factors are important for the optimization of existing public spaces for a better running climate.

To give better insight in the policy context of this research, we start by outlining the context and purpose of this study. Building on these, we describe the approach we used to get

insight in how user generated running data could be relevant for the urban design of public spaces. Through a combination of data visualization techniques and exploratory data analysis we present a series of iterations that gave us insight in the value of user-generated running data for urban planning. Through three iterations we show how we have been looking for factors that influence the quality of running environments and how this can be valuable for policy making of public places. We approach this from an urban design perspective and involve big data to research its potential benefit for this discipline. Finally, we discuss the qualities of running-app data with regard to designing for urban spaces and outline future steps for this research.

Policy and Research Context

The research is situated in and focuses on the city of Eindhoven, located in the south of the Netherlands. Eindhoven is the fifth-largest city in the Netherlands, with a population of about 223,000. Eindhoven pursues to position itself (internationally) as an innovative and sports minded city, focusing both on attracting elite sports and on providing ample good quality sport facilities and public areas for leisure time sports (City of Eindhoven, 2008).

The city government largely controls the development of sport-infrastructure itself. Three areas have been appointed with a specific focus on the provision of mass sport and physical activity opportunities. Each of this areas has its own management which guards the integration of nature, sport, art, culture, education and recreation. The three sport areas are to be connected by the 'Green Y', a Y-shaped combination of natural green areas in Eindhoven. (City of Eindhoven, 2008)

The further development of the provision of mass sport and physical activity opportunities in Genneper Parken is the research context for this study. Over the last 30 years, this public park, one of the three areas of interest for the city of Eindhoven, has been developed into a recreational sports park that is well balanced with other urban characteristic (e.g, ecological structure, original landscape, historic village) (Genneper Parken, 2016). The focus on sport and recreation was already mentioned in local development plans in 1988 (City of Eindhoven, 2004). Since then, multiple local strategy planning documents have confirmed and maintained this focus (Weijschedé & de Poel, 2002, City of Eindhoven, 2004, Province of North Brabant, 2006). Evidently this is an area that received extensive attention with regards to sports and upon observation we clearly recognize these qualities.

The city of Eindhoven considers participation in recreational running as an important goal in their sport policy. Indeed, the promotion of running in Genneper Parken is considered to be a powerful tool to stimulate participation in sport and physical activity in Eindhoven.

In the present paper we use user-generated big (running) data to unravel which qualities make up for a good running environment and what other implications access to this data could have for urban planning of public spaces.

Method

To investigate the role of user-generated big running data, this study utilized data from two popular running apps (for less-experienced runners) in the Netherlands and Belgium. One of the advantages of running related smartphone applications is their ability to track behaviours over time in the daily urban environment (Vos et al., 2016). Start2Run (STR) [(Start to Run, 2016) and Hardlopen met Evy (HME) (Hardlopen met Evy, 2016) (translated Running with Evy) are running apps that intend to motivate people to start running by providing training schedules and feedback and offer them insight into their running patterns. In essence both running apps are identical, as they share the same owner EnergyLab (EnergyLabs, 2016). STR is positioned for the Belgian market and HME for the Dutch market. Both apps are only available in Dutch.

The data collected from the app consisted of full GPS trails for each run (1.5 million runs in total). On top of these run trials a metadata set summarizing these runs is available. This summary data includes a run- and user-id, timestamp at the start of run, duration, distance, average speed, effective time (duration from start to stop minus the pause time) and training id (if a specific training session provided by the app was followed).

To explore how this data can be relevant for urban planning and design of public environments, an iterative approach was applied. In a *first iteration* (interactive) data-visualization techniques were used to get grip on the quality of our data and to understand basic geographic characteristics of runs. Besides taking a geographic (collective) perspective, a side step was made to exploring what insights an individual runner's perspective may bring us. Based on insights from the first iteration, the *second iteration* revolved around characteristics of quality of urban running environments. A mixed-data source approach was used to analyse a selection of twenty running locations, based on city level visualizations. Both running-app data and qualitative checklist-data were combined. Via the qualitative checklist, a set of variables was scored for each of the locations. These variables were: possible running distances, running surface, background soundscapes, green or natural environment, (artificial) light, accessibility, signposting and state of maintenance. In a third iteration we tried to scale this further by comparing 271 handpicked running spots in the Netherlands (NL) and Belgium (BE) to generate a set of focus features as input for a future extended and more in-depth mixed method approach.

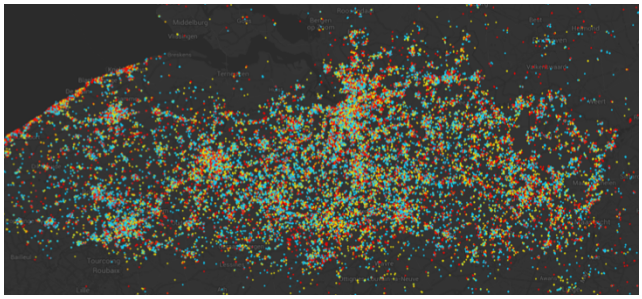

Figure 1: Heat map of Belgium, colored by daily time clusters.

Figure 2: Heat map of Gent and Antwerp

Figure 3: Visualizing different variables simultaneously in Eindhoven. (seasons)

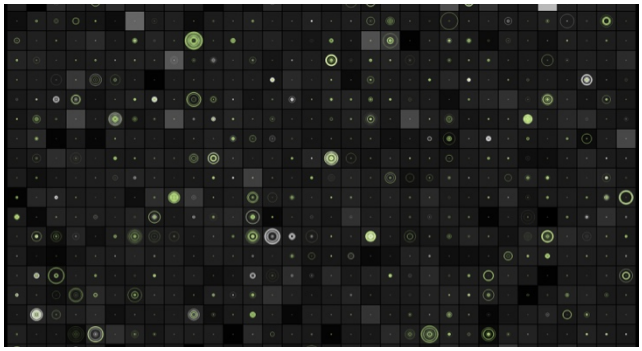

Figure 4: Visualisation of running routines per user.

Results

In total the dataset contains 1,490,145 runs since 2012. Exactly 40% of these runs is created by the Belgian version of the app (STR). Of all runs only 76936 (5.2 %) is located outside NL and BE, forwarding the clear geographical focus. Overall runs have an average speed of 8.0 kilometres per hour with an average distance of 3.9 kilometres. Compared to the Amsterdam data of the popular app Strava (Strava Insights, 2016) this is 23.5 % slower (10.45 km/h) and 53% less far (8.2 km). This is a clear indicating these apps focus on less experiences runners; a different target group. As we approach this from an urban planning perspective we first explored the data to get insight in the geographical distribution of the runs. Outlined below is a description of activities of the three different iterations, combined with insights that connect these iterations.

<u>Iteration 1: Data Visualization</u>

Based on the starting locations of each run, a geographical overview was created (Figure 1). Next to the aforementioned available metadata of these runs (e.g., start time, duration), each run was given a number of additional attributes (i.e. daily time cluster (morning, afternoon, evening), the day of the week and the month it was ran). Comparing these location based starting point visualizations, that show these characteristics (either color-coded or only showing fitting runs), provided a first insight in running behaviour that varies over time. Differences in quantity between varying types of runs were clearly visible (e.g., more summer than winter runs), while the relative geographical spread was hardly affected by these distinctions. To have more detailed insight in how this changed on a city level, we merged the full GPS trails with the metadata to create city level heat maps that showed all running trails in the selected cities, instead of just the starting position (Figure 2). Again, extra attributes were calculated and added to the dataset. (i.e., season, week or weekend, daylight or dark based on sunrise and sunset times). This made it possible to visualize the GPS trails combined with the running metadata (Figure 3). These heat maps were created for five different cities in NL (4) and BE (1). Based on the runs located in these areas, metadata was calculated and added to these maps (i.e. total distance, amount of runs, unique users, average speed).

Next to these location centric visualizations, user centric visualizations were created. Figure 4 visualizes the running routines of 750 (random) users. One square is dedicated to each user. The circles represent a run, the distance between the circles is depending on the amount of time in between these runs (like the year-rings of a tree). When a training exercise from the app is followed the circle is coloured green, otherwise it is coloured white. These simple visualizations enabled us to quickly visually compare the running routines of different users, to see the time interval

between runs or structured routines. These individual patterns added additional information which can be relevant the urban planning of public spaces.
The running heat maps of all cities clearly highlight green and natural environments. Comparing daylight and night-time runs, reveals significant differences in running patterns. Indeed, these visualizations show that most of these 'green' locations lose their attraction after nightfall. Of course, lack of artificial lights and safety will be a defining factor when explaining this difference, but there may well be others. Where during the daytime the parks are the clear hotspots in the city, at night these hotspots move to streets in residential areas (Figure 5). This contrast can have a variety of possible causes. It could mean that people choose to run closer to home later in the evening, that the 'green' environment loses its attraction when it's dark out or that the difference in social control and 'feeling safe' between a night-time park and a night-time residential neighbourhood is important to runners. Even though the data might tell us something about how these spaces are used by runners, it contains little information about why people act this way.

Figure 5: Heat maps of Eindhoven region before sunset (left) and after (right)

The user centric visualizations give a slightly better insight in personal motives as it differentiates individual from the group. Keeping some of these individual run characteristics in mind we decide to focus our next iteration on the location based approach as our prime interest is in the environmental characteristics.

Iteration 2: Mixed Data Source Approach
The first iteration, using heat maps and running statistics from the database, showed a clear difference between the use of parks and other natural areas and more urban running sites in (and around) the city. Finding out exactly which environmental factors play a part in causing this distinction would be a great step towards determining what makes a good running environment and thus provide valuable insights for urban planning for active and healthy public spaces. This second iteration therefore had a mixed data source approach, focussing on collecting environmental characteristics of twenty running locations and comparing these with the running data from the app.

The city level visualizations were used to select interesting running spots in the cities. For these areas more detailed and zoomed in heat maps were created, again accompanied by area-specific statistics. To determine if runs classified as being on a certain track, we used a two-step approach. All running locations were described by a tightly fitting bounding box. Runs that had a starting location less than 10 kilometres away from this bounding box were selected. For these runs it was verified whether one of their coordinates would fall inside the inner bounding box.
The tracks were qualitatively and manually scored on environmental characteristics, using a simple checklist to keep the observations as objective as possible. Through statistical analysis, we then combined this qualitative scoring with the collected GPS trails to find factors for optimal running environments. Based on Pearson linear correlation coefficient calculations, there are a few interesting potential influencers. A correlation was found between the number of runs on a track and the average speed there. Also, the maximum distance of a track (being the maximum distance that can be run without repetition) appears to influence not only the number of runs, (and with that the total distance run) but also the average distance.

Although some indications of relevant factors came forward from this iteration, they were not so distinct that a total of twenty scored tracks was enough to draw solid conclusions. An initial insight in future focus-areas is gained but additional data, especially concerning the environmental characteristics and preferably also qualitative experience data, will have to be collected to really understand and break down environmental influence on running behaviour.

Iteration 3: Running Location Analytics
Iteration 2 clearly highlighted the need for more contextual and experiential data to understand which qualities are influencers of good running environments. This iteration therefore sets out to get further insight into environmental elements that could be influential by expanding the set of evaluated locations and reducing the number of scored attributes.

In this study we are interested in factors that can be influenced by urban design. For instance, if a longer distance track turns out to attract more runners this is something we can influence. If people only run in good weather this becomes harder to influence; unless influenceable factors are at the heart of this (e.g., slippery or muddy surfaces). To gain more insight into these factors we manually marked 271 running tracks; 87 city parks and 184 'Finse Pistes'. A 'Finse Piste' (bark running track) is a man-made, looped running track, covered with a soft top layer of woodchips or tree bark. (Borgers et al 2016) The city parks and 'Finse Pistes' have different characteristics and within both categories locations also greatly vary.

Based on the full GPS trails we clustered runs to the marked locations. All ‘Finse Pistes’ are looped. To determine which runs took place on such a track only runs coming within a 200-meter radius of the centre coordinate were added to the set. Instead of pre-defining factors we hoped to find outliers in running tracks. In a next step we could then qualitative analyse running tracks across a spectrum to define new factors to score.

The shown correlation diagram (Figure 6) is a first experiment to see how variables in our set are related. There are expected moderate correlations between variables, for example the time of the day (i.e., morning, afternoon, evening) and whether it's a Monday to Friday or a weekend run.

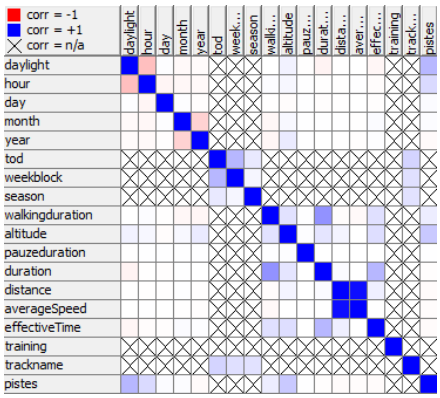

Figure 6: Correlation diagram of environmental attributes.

Comparing the ‘Finse Pistes’ with the city park tracks, we note that on average app users run 0.59 km/h slower on the ‘Finse Pistes’ than city parks. As we know these locations to be focussing on novice runners and have a different surface, this matches our expectations.

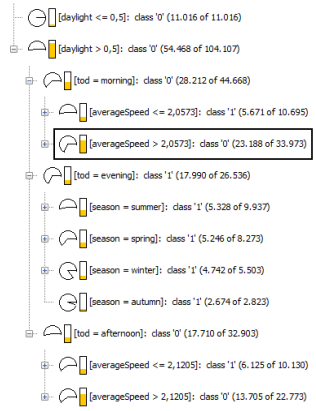

Figure 7: Outcomes initial prediction model

Next, they were compared in an initial prediction model (decision tree learner & prediction (Figure 7)) to investigate if there are strong differences in characteristics between locations. The purpose of this was to see if the track could be predicted based on the other attributes. Aside from a strong predictor that later turned out to come from incorrect and therefore misrepresentative daylight variable, this experiment showed little strongly predictive values. Future repetition of this experiment with correct daylight information could therefore potentially still show predictive value for this attribute.

Discussion and Implications

This study explored the possibilities of user-generated big data for the urban planning of active and healthy environments, and providing recommendations for physical activity policy making. The results show that from this data green and natural environments in cities can clearly be identified as running hotspots, showing the attraction of these areas on runners. A notable difference between daylight and night-time runs however, shows that this attraction only lasts during the day. These findings are in line with the findings of Borgers et al. (2016) who show that a good light setting is a primary condition for people to value running tracks. Additionally, interesting potential environmental influencers for the attraction of green areas for runners are found, including the maximum uninterrupted length of a track.

To make user generated big data relevant for urban planning additional information about contextual and experiential factors is needed next to the GPS and meta data from the running app. The iterative approach of data visualization and analytics proved an effective way of zooming in on the data and it's promising attributes. As a first step in the process of determining important environmental factors for an optimal running climate, in urban areas, it provides first insights and a broad foundation for further research.

This data used for this study has some limitations which have to be pointed out. Since the data was collected using a specific running app, we must note that not every runner is using this app, and the users are at most roughly evenly spread out over the inhabitants of Netherlands and Belgium, never exactly. App users might not always take their phone when going for a run, so there could actually be more runs than registered. Finally, this app targets novice runners and is therefore likely to have a higher percentage of starters than the actual running population in the Netherlands and Belgium. Because of the amount of data collected, however, it still gives a realistic insight in societies running behaviour and preference for running locations.
Contrary to the large quantity of running data collected from the apps, the number of qualitative data collected about the running environment at the chosen locations

proved insufficient for a credible conclusion. For this data to provide real insight in the exact aspects of public space that attract or repel runners, more running spots would have to be scored. This issue came from the absence of environmental information in the app data. In order to determine exactly what environmental factors influence running behaviour (and maybe even to which extent), the dataset needs attributes concerning that environment. As mentioned before, this data was manually collected later, on a much smaller scale, while including a short questionnaire or even just a location-score element in the app could generate lots of these data, provided by the runners themselves.

The use of squares around running locations to determine the amount of runs there was good for this rough first insight. It is suggested that in future research the location boundaries will have to be followed more closely to be able to give exact data. In this study the focus was on different city parks. A run would be connected to a park when one of its running locations would fall inside the bounding box of that location. In the first iteration we showed a more user-centric approach. In the future it would be of interest to focus on how individuals behave differently in different environments. It would be interesting to research what deviations from structured routines are caused by, and if environmental factors play a role in this. If different runs have a different path or terrain, what is the difference between these runs (e.g., do people mostly run longer, faster or more uninterrupted in parks?) This way it could be possible to exclude more unknown factors from our models to better understand the nuanced impact of specific environments.

Making visual representations of the data made them instantly insightful for the research-team, and proved to be very useful when presenting findings and proceedings to the other involved parties. It also turned out to be an effective way of checking whether there were 'strange things' happening in the file-writing process, and runs with incorrect data could often easily be spotted.
In a more in-depth study of this data and or topic there will also have to be corrected for several factors, e.g. for the size of the running location; the amount of inhabitants of the city; amount of nearby residents with some of their characteristics, and the number of runs per park.

Although the prediction model from the third iteration was a small exploration to find qualities of running environments on a larger scale with limited concrete insights, we see clear opportunities for continuing this step in the future. In the last examples we tried to predict which park a run situated in, based on a set of variables. This predictor aims to do this for all parks in general, not for specific parks. Focussing further on detailed scenarios might give us more insight in which factors influence good running environments.
During the study, the research topic proved much broader than initially thought. Every iteration, new insights raised new questions. This made the process both very interesting and somewhat unpredictable.
Especially running statistical analysis was challenging, because instead of trying to find relations between given elements, where we were actually looking for the elements that might have a relation, which were not necessarily present in the dataset yet. On top of this we were primarily looking for elements that could be influenced from our urban design perspective. The combination between qualitative analysis through visiting parks, looking at running routes at different locations and data analytics proved a value approach to advance in this.

In the 'Policy and Research Context' section we described the background of this study and our initial interest in the Genneper Parken, as one of the focus areas for urban activity of the city of Eindhoven. An already well developed area, with regards to sports and activity, which challenged (and keeps challenging) our research to use running data to find sweet spots for improvement. Throughout our analysis we have seen clear indications for the importance of running environments that provide uninterrupted running experiences. By using the user-generated running data, we aim to better understand how this contributes to a good running environment. However, we noticed that larger parks are the host of longer runs and we have seen that longer runs have a higher average speed (i.e. more experienced runners). Although the three active environments in Eindhoven are geographically connected in a Y shape, this is not necessarily visible from our heat maps. Through small but innovative solutions, as green zones for runners (Megens et al., 2016), we hope to be able to better connect these areas to shape a better running environment. While we further investigate how this data can help us in better understanding these running environments, a long-term partnership with the city of Eindhoven has been established to further analyse user generated big data, in combination with survey data and observations, to support their policies for active and healthy public space.

Acknowledgements
We would like to thank Energy Lab for making their running-app data available for this study.

References

Bauman, A. E., Reis, R. S., Sallis, J. F., Wells, J. C., Loos, R. J., Martin, B. W., & Lancet Physical Activity Series Working Group. (2012). Correlates of physical activity: why are some people physically active and others not?. *The lancet*, *380*(9838), 258-271.

Blair, S.N. (2009). *Physical inactivity: the biggest public health problem of the 21st century.* British Journal of Sports Medicine, 43, 1-2.

Borgers, J., Vanreusel, B., Vos, S., Forsberg, P. & Scheerder, J. (2016). *Do light sport facilities foster sports participation? A case study on the use of bark running tracks.* International Journal of Sport Policy and Politics, 8(2), 287-304

Breedveld, K., Scheerder, J., & Borgers, J. (2015). Running across Europe: The way forward. In: J. Scheerder, K. Breedveld, & J. Borgers (Eds.), *Running across Europe. The rise and size of one of the largest sport markets.* (pp. 241-264). Basingstoke: Palgrave Macmillan.

City of Eindhoven. (2004). *Structuurvisie Genneper Parken*. Eindhoven: City of Eindhoven.

City of Eindhoven. (2008). *Sportnota 'Hé, ga je mee?'* Eindhoven: City of Eindhoven.

Cleland, V. J., Ball, K., King, A. C., & Crawford, D. (2010). Do the individual, social, and environmental correlates of physical activity differ between urban and rural women?. *Environment and behavior*, 0013916510393275.

Energy Lab. (n.d.). Retrieved August 14, 2016, from http://www.energylab.be/

Frumkin, H. (2003). *Healthy Places: Exploring the Evidence.* American Journal of Public Health.

Genneper Parken. (n.d.). Retrieved August 14, 2016, from http://genneperparken.nl/algemeen/de-totstandkoming-van-genneper-parken.html

Granner, M. L., Sharpe, P. A., Hutto, B. E., Wilcox, S., & Addy, C. L. (2007). Perceived individual, social, and environmental factors for physical activity and walking. *Journal of physical activity and health*, *4*(3), 278.

Hardlopen met Evy. Je persoonlijke coach en looppartner op je smartphone! (n.d.). Retrieved August 14, 2016, from http://www.hardlopenmetevy.com/

Larson, T. A., Normand, M. P., Morley, A. J., & Hustyi, K. M. (2014). The role of the physical environment in promoting physical activity in children across different group compositions. *Behavior modification*, *38*(6), 837-851

Mansfield, E. D., Ducharme, N., & Koski, K. G. (2012). Individual, social and environmental factors influencing physical activity levels and behaviours of multiethnic socio-economically disadvantaged urban mothers in Canada: A mixed methods approach. *International Journal of Behavioral Nutrition and Physical Activity*, *9*(1), 1.

Megens, C., Neutelings, I., van Hout, B., Janssen, M., & Vos, S. (2016). Run!, Developing a high tech running path in the park. Science and Engineering Conference on Sports Innovations (SECSI) 2016, April 8, Amsterdam (The Netherlands).

Province of North Brabant (2006). *Natuurgebiedsplan 'Dommeldal-Zuidoost' - Streefbeelden en subsidies voor natuur en landschap*

Sallis, J.F. et al. (2016). *Physical activity in relation to urban environments in 14 cities worldwide: a cross-sectional study.* The Lancet April 2016.

Start to Run. Je persoonlijke coach en looppartner op je smartphone! (n.d.). Retrieved August 14, 2016, from http://www.start2run.be/

Strava Insights – Explore the world through Strava data. (n.d.). Retrieved August 14, 2016, from http://insights.strava.com/en-us/worldwide/?unit=imperial

Vos, S. (2016). *Designerly solutions for vital people.* Eindhoven: TU/e.

Vos, S., Janssen, M.A., Goudsmit, J., Lauwerijssen, C. & Brombacher, A.C. (2016). *From problem to solution: Developing a personalized smartphone application for recreational runners following a three-step design approach.* Procedia Engineering, 147, 799-805.

Weijschedé, T.J. & K.R. de Poel, 2002. *Genneper Parken in beeld. Een cultuurhistorisch en landschappelijke verkenning ten behoeve van de planvorming voor Genneper Parken in Eindhoven.* Wageningen, Alterra, Research Instituut voor de Groene Ruimte. Alterra-rapport 588.

World Health Organization (2010). *Global Recommendations on Physical Activity for Health.*

World Health Organization, (2010). *World health statistics 2010. World Health Organization.*