Open Data analysis to retrieve sensitive information regarding national-centric critical infrastructures

Research Project 2

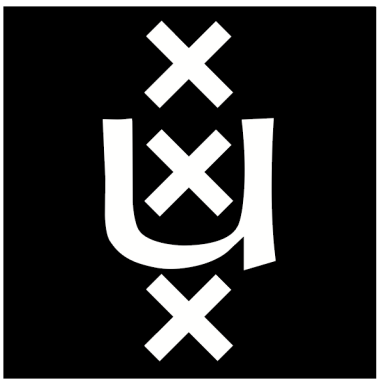
Renato Fontana
renato.fontana@os3.nl

Supervised by:

Ralph Dolmans, NLnet Labs

Benno Overeinder, NLnet Labs

University of Amsterdam
February 3, 2014



Abstract

Open Data repositories store a variety of information from country governments and private sectors that can be used to conduct research, development and analysis. The volume of data made publicly available supporting the Freedom of Information Act is increasing exponentially as new data is being published by government and private sectors upon users request. A concern with the releasing of government and private sector data is that sensitive information can be obtained by means of visualization techiques. The project presented in this report uses a method of visual analytics known as feedback loop process to acquire insights from data. Visualizations were created to identify patterns in areas with a high number of Critical Infrastructures sector. The results show that its possible to retrieve precise locations where critical infrastructures overlaps by applying visual analysis techniques.

Keywords: Open Data; Public Data; Sensitive; Critical Infrastructures;

Contents

1 Introduction

Open data is data that can be freely used, reused and redistributed by anyone – subject only, at most, to the requirement to attribute and sharealike [2]. From one perspective, the idea of a democratization of data arises from the desire to obtain further insights from a variety of sources. For example, government accountability, companies acceptance, financing statistics, national demographics, geographic information, health quality, crime rates or infrastructure measurements.

Advanced Open Data analysis can provide government, private sector and citizens with enhanced knowledge gained through the development of new applications. Such applications assist on information representation in alternative ways. Therefore, interested users making use of application supported by Open Data knowledge are capable of obtaining further insights complementary to standard news from traditional media communication channels. The output from these tools can be enhanced visualizations, interactive graphs, illustration of most relevant facts, or simply a matter of showing data in a more user-friendly way.

The facts presented so far contributes to a democratic scenario where sharing and collaborating is desirable by all means. Nothing is referred to the level in which data should be open and available for others; apart from the US Open Data website which specifies rules that public information must not contain National Security information.[1] Another perspective, with a more critical point of view, such statement regarding Open Data may cause unrest in Government agencies that seeks to safeguard national sensitive information. A high volume of information made available without proper accordance with its classification can be used as a strategic source of knowledge by third parties or even terrorists.

Gradually, private sectors and country governments are releasing several knowledge bases to public access. The current scenario of publishing data seems to be a competition among countries to be on top as the one with most information made public, as it can be seen at Open Knowledge Foundation Country Index [2]. This rapid release of data can pass unnoticed in the information classification criteria and therefore increasing the likelihood to the disclosure of sensitive information.

[1] http://www.data.gov/data-policy

The disclosure of a single dataset may not represent a security risk, but when compiled with further information it can truly reveal particular Critical Infrastructure areas. This public data is mainly reffered as 'sensitive but unclassified' information [3].
Critical Infrastructures are essential services that contribute to the stability and security of a Country [15]. A high volume of data currently available in Open Data repositories can be mapped to Critical Infrastructure, but this data may not appear harmful on its own. Additional datasets can be indirectly obtained by means of Freedom of Information Act request. The FOI Act differs in its reasoning according to each Country's perspective but its general understanding can be interpreted as follows in the US Department Homeland Security FOI Act report:

"The FOIA establishes for any person—corporate or individual, regardless of nationality—presumptive access to existing, unpublished agency records on any topic"[12].

The concept of what is a Critical Infrastructure must be adequate to specific scenarios. This means that what is critical for one Nation may not be seen as a high risk to another. Therefore, proper classification and scope must be taken into account when performing research in such field. Such classification must also take into account the chain of Criticial Infrastructues cascading failures [11].
Table 1 contains all categories and sectors regarding infrastructures classified as critical by the Dutch Government [1].

This project's aim is to visually represent national-centric Critical Infrastructure areas by taking into account public information available in Open Data repositores. The research is scoped to Dutch Critical Infrastructures, therefore it is also based on existing publications regarding this subject as input to categorize each sector and its interdependencies. As classified by the Dutch Government, Critical infrastructure includes the business enterprises and public bodies that provide the goods and services essential for the day-to-day lives of most people in the Netherlands [1]. Also under the Dutch Government resilience approach, the concern with Nation Security provides more details about existing strategies to prevent disaster or crisis.

Category	Critical Sector
Energy	Electricity, natural gas and oil
Telecommunications and ICT	landline and mobile telephony, radio, broadcasting and the internet
Drinking water	The water supply
Food	the food supply (including in supermarkets), food safety
Health	emergency and hospital care, medicines, vaccines
Financial sector	payments and money transfers by public bodies
Surface water management	water quality and quantity (control and management)
Public order and safety	public order and safety
Legal order	the courts and prisons; law enforcement
Public administration	diplomacy, public information, the armed forces, decision-making
Transport	Amsterdam Schiphol Airport, the port of Rotterdam, highways, waterways, railways
The chemical and nuclear industries	the transport, storage, production, and processing of materials

Table 1: Dutch Critical Infrastructure sectors

1.1 Research Question

Information made freely available, supporting the idea of democratization of data, might introduce an undesired scenario that most Countries and Governments would prefer to avoid. Depending on the disclosure degree of information in Open Data repositories, malicious users may obtain knowledge to categorize national-centric Critical Infrastructures.
The goal of this project can be summarized in the following questions:

- **Can users make use of Open Data databases to retrieve country sensitive information?**
- **How can this information be visually represented in a way that allows easy identification of critical and strategically important areas on detailed level?**

For proof of concept reasons, research questions are limited to Dutch Critical Infrastructures. Therefore, Netherlands and particular areas are taken into account during analysis.

1.2 Outline

This report is structured as follows. Section 2 presents an overview of existing studies performed that are to some extent related to this research's approach. Section 3 elaborates on the approach used to obtain insights by using a feedback loop process technique. Visualization methods for pattern recognition are described in Section 4. Section 5 elaborates on data analysis and results from the visualization system. Conclusions are drawn based on the final results in Section 6 and Future work in Section 7.

2 Related Work

To our best knowledge, there is no existing research based on Open Data that pinpoint precise areas with a collection of Critical Infrastructures. There are related studies that focus on the risk of public data availability on Critical Infrastructure Protection. In Abbas et al., 2006 a study refers to public data as "sensitive but unclassified" information [3]. That is, information that may not on its own appear harmful, but when compiled with other data can be truly revealing about an individual or critical infrastructure. Such study concludes on the risk of sensitive data being available in the public area and suggests the increase of awareness in Australia's scenario.

A report for Congress in the US highlights the concerns in information sharing and the balance between public's right to know [12]. One of the proposed measures was to exempt Critical Infrastructure information from disclosure under the Freedom of Information Act.

The definition of Critical refers to an entity that is essential or vital in nature [6]. For Critical Infrastructure the definition follows to what are the essential services that contribute to the stability and security of a country [15].

Most existing research relates in some extent to this paper by addressing the nature of Critical Infrastructure and its modeling approach to dependencies. In Rinaldi et al., 2001 a study regarding the Critical Infrastructures is conducted to show a chain of interdependencies among each of the critical resources and the impact of these in case of failures [15]. One of the highlights in this study states:

"What happens to one infrastructure can directly and indirectly affect other infrastructures, impact large geographic regions, and send ripples throughout the national and global economy."

It also takes into consideration the nature of data availability and its granularity. The amount of data is a trade-off with model and simulation fidelity. One of the concerns about the availability of this data is related to security and proprietary issues.

"A highly detailed, comprehensive database of national infrastructures would be a valuable target for hackers, terrorists, and foreign intelligence services particularly if it were coupled to advanced modeling and simulation tools." The above lines shows the substantial need to balance the security and the need to make data available for researchers.

Luiijf et al., 2009 [11] explored a method focused on public reports of Critical Instrastructures disruptions, collected from public sources like newspapers and internet news outlets. Analysis classified events by cascading effects (initiating, resulting and independent events) which suggested that cascades failures are fairly frequent.

An early research on "Mapping the Dutch Critical IP Infrastructure" [4] showed the level of insights that is possible to be obtained from public sources. Even without any privileged access to information, and based on specific assumptions, it was possible to observe that most Dutch Critical Infrastructures organizations rely on foreign communication providers.

The present reasearch in our project extends to the field of Critical Infrastructures identification by taking into account the nature of published content by Government and private companies under the Open Data initiative. It also ties together the risk of existing public information to Critical Infrastructure Protecture measurements.

3 Approach

As stated in Section 1, this paper is scoped a scenario only considering The Netherlands as proof of concept. This section follows with approaches used for data acquisition, data sanitization, initial analysis and the method used to create visualization derived from visualization insights.

In order to answer the research questions, the methodology used in this paper follows the principles of hypothesis development and experimentation. The objective is to formulate hypothesis and visualizations based on data, and the outcome is to obtain insights from particular facts by aggregating them into multiple Critical Infrastructure levels, described later as layers. Such method is explored in Visual Analytics studies and presented by Keim as a feedback loop process, or the Visual Analytics mantra [10].

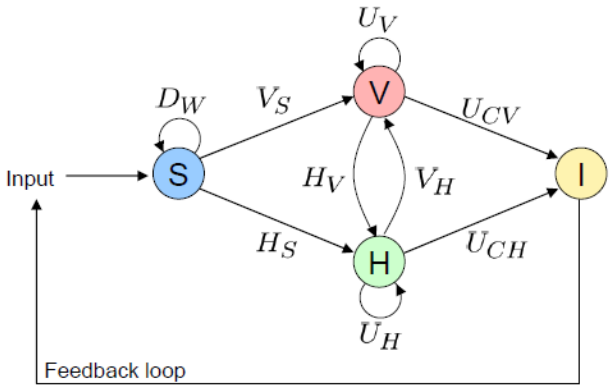


Figure 1: Visual Analytics Process as described in Keim at al.[16, p.5]

As shown in Figure 1, the process is a continuous interaction between Visualization V and Hypothesis H based on the existing Data Source S. The outcome is to obtain Insights I, which can be confirmed or invalidated based on H_V and V_H interactions.

For each dataset containing relevant Critical Infrastructure information, a visualization layer is placed under that category. All layers are combined into a visualization system which supports the researcher to quickly recognize patterns in images.

Features to scale down the visualization are also explored to provided a finer granularity of the data. At first, a general overview is presented, then the features to scale to different levels. At each level, visualizations are capa-

ble to provide details from each identified Critical Infrastructure layer. The outcome is high dynamic and configurable visualization of data.
Such interactive technique is defined in the "Visual Information-Seeking Mantra" defined by Shneiderman [16]:

"Overview first, zoom and filter, then details-on-demand"

4 Visualization of Open Data

The optimal solution to quickly gain insights, is to integrate the multiple Critical Infrastructure dataset layers into a single visualization. At first, a data acquisition phase was necessary to fetch content from open data repositories. Second, a set of the most relevant datasets were selected to the sanitation phase. The third step focuses in generating visual representation of data to identify patterns. The final feature of the system is to provide the researcher with filtering capabilities, by selecting layers representing each Critical Infrastructure sector, to enhance pattern recognition. Each step is described in details in the following subsections.

4.1 Data Acquisition

According to the Open Data Partnership,[2] at least 58 international countries participate on the democratization of public sector information. At the time of this research, The Netherlands holds the 5th place in the score rank of most content made public. Over the Open Knowledge Foundation Index [3], it is possible to retrieve the very first links containing datasets from Netherlands and downwards. These links can change in the future, but they mainly refer to reliable sources to obtain public data.
It is also possible to find a wide range of non-listed repositories holding information that could be translated to Critical Infrastructure identification areas. Full platforms are available to users to start their own open-source data portal platform and store data.[4] Therefore, this wide range of available sources introduces a high effort in obtaining data publicly without discarding any relevant information.

The data acquisition phase followed the same structure of Critical Infrastructures sectors provided by the Dutch Government as guidance to fetch

[2]www.opengovpartnership.org/
[3]Open Data - https://index.okfn.org/country/
[4]http://ckan.org

datasets. See Table 5 on Appendix.
Techniques to fetch data covered the use of web crawler scripts, visual inspection, advance filtering in search engines using keywords and manually downloading of contents.

The outcome of this stage is the identification of 2 to 5 stars sources, as is classified by Tim Berners-Lee on the Open Data develpment scheme.[5] This classification attributes a star grade based on the content available on the Open Data repository or application. One star for .pdf documents and static images. Two, for structure data in spreadsheets. Three, for the use of non-proprietary formats e.g CSV. Four for the use of URIs (XML/JSON). And 5 stars for data linked to other data contents.
Therefore, for the data acquisition phase, scope is on content types in .xls, .csv, URIs links and sources with aggregated data available for download. Webpages from governments, private companies and self-maintained CKAN repositories supporting the Open Data initiative were targets of this data acquisition process. The mapping for each sector to a know open data repository URL can be found in Appendix.

Most of the identified URLs contain data that overlaps, this occurs due to the close relationship and references present in those repositories. It also suggests the re-host or mirroring of content to other repositories. At this phase no datasets are filtered out from analysis.

4.2 Data sanitation and initial selection

Data obtained from the earlier phase resulted in a wide range of content in different formats and structures. To address this issue, sanitation techniques were applied to create a more coherent data structure. Automated solutions were developed to remove arbitrary, unstructured and blank entries from the datasets. This phase provided a more reliable data source to proceed with the upcoming analysis. Only datasets with potential information were subject to sanitation, further data with no reference headers, no documentation or unrelated to other sources were filtered out of the analysis phase. Due to the limited research time frame, selecting specific datasets is also a measure to scope this research project.

The initial visual inspection reveals that most datasets contain common identifiers which reflect to geolocalization points in different metrics. Namely,

[5] http://5stardata.info/

headers containing:

```
FullAddress, PostCode, COORDS, LAT, LONG, City,
CENTERLAT, CENTERLNG, Locatie, adres, buurt, wijk, gem.
```

At this point, it is possible to gain initial insights that most fetched datasets establish some references to specific country areas in its public nature.

Such geolocalization entries were combined with filetype .shp [8] which were also obtained during the data acquisition phase. The *Centraal Bureau voor de Statistiek*(CBS) holds an open base for shape files with different administrational levels boundaries; namely, neighborhood, district and city. Such files were parsed using open source geographic information system, QGIS [13] and converted to .kml [14] format using existing conversion tools.[6]

The system was developed on top of Google Maps API and Google Fusion Tables to dynamically integrate geocoding features. This setup greatly increases map visualizations by identify datasets fields containing localization referenced entries. Figures 2 and 3 illustrates the two types of the systems visualizations layer with all communication antennas located in The Netherlands.

Figure 2: Dots

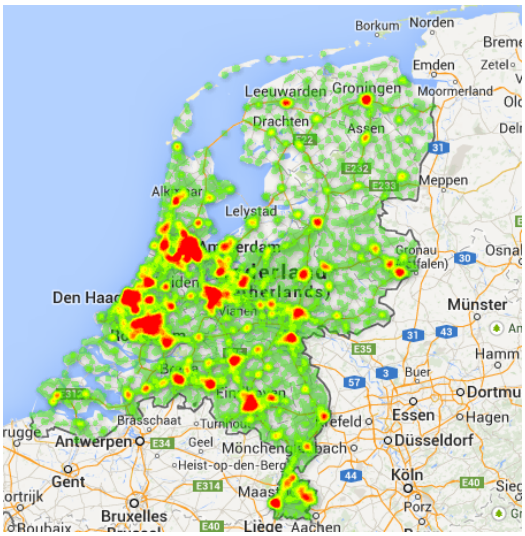


Figure 3: Heatmap

[6]http://www.shpescape.com/

In Figure 2 each telecommunication antenna is illustrated by a single dot on the map representing its location. Another type of visualization in Figure 3 uses a Heatmap visualization to illustrate areas with a higher number of antennas. Hot colors for areas with a large grid and cold colors for smallers.

Following this interactive visualization system and configuring each one of the available layers, the researcher is able to use human cognitive reasoning to identify patterns, gain insights and derive conclusions.

5 Data Analysis and Results

For each dataset obtained from the earlier phases of data acquisiton and data sanitation, a visualization layer is created to represent that specific Critical Infrastructure. The collection of visualization result in the amount of 22 interactive layers showing statistics and values for regions or specific points on the map. The Open Data sources used for each of these layers is provided in the Appendix.

The gathered data reflects to a snapshot of the current volume of public data that translates to CI in the Netherlands. Many other sources making use of live APIs contained real-time information. It was not possible to integrate such data due to discrepancies in relative time with other datasets. E.g. the latest energy consumption dataset from 2011 combined with real-time transportation metrics, does not provide a reliable comparison measurement.

Analysis follows with the first visualizations on the macro level in attempt to answer the initial research question. Therefore, specific layers at country level (Netherlands) are selected to create hypothesis and gain insights. For the sake of the project scope, only a section of the available datasets related to critical infrastructures were analyzed in a detailed level. All other layers were implemented and are visually available over the visualization system. No further analysis were executed to investigate those due to time limitations for the research.

5.1 Nation level analysis

As an initial approach to identify patterns, the visualization is centralized at the Nation level Netherlands. The base hypothesis in this case is to identify if there is a high number of Critical Infrastructure events combined within a specific region. Analysis in this level considered the following dataset layers:

- Communication Antennas location;

- Power Plants location;
- DataCenters - Internet Exchange peerings;

Figure 3 in Section 4.2 illustrates a heatmap representation of areas with the most numbers of communication antennas supporting GSM 1800, GSM 900, LTE and UMTS technology. It is possible to identify warm areas in Den Haag, Rotterdam and Amsterdam with a slightly big grid.

The second layer shows in a heatmap style areas with higher number of Power Plants illustrated by Figure 4.

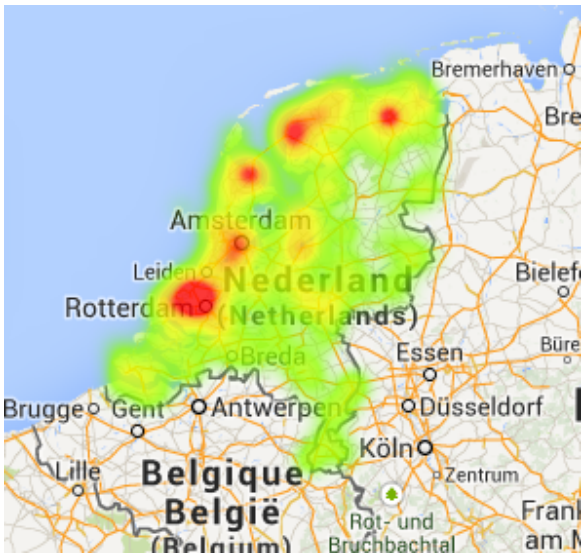


Figure 4: Power Plants - Heatmap

It is possible to identify that surrounding areas of Den Haag, Rotterdam and Amsterdam once again provide higher numbers related to these two layers of Critical Infrastructures. At this point, visualization can be zoomed in to a lower level to gain further insights.

To add the 3rd layer with Data Centers location, the visualization scales down to a lower level to highlight the identified cities with high occurrence of Critical Infrastructure spots.

Figure 5 illustrate the combination of Power Plants with red dots, Data Centers with blue dots and an underlying heatmap with the communication antennas grid. As far as analyzed, insights reveals that Amsterdam hold

the most identified critical infrastructures related specifically to the current configured layers.

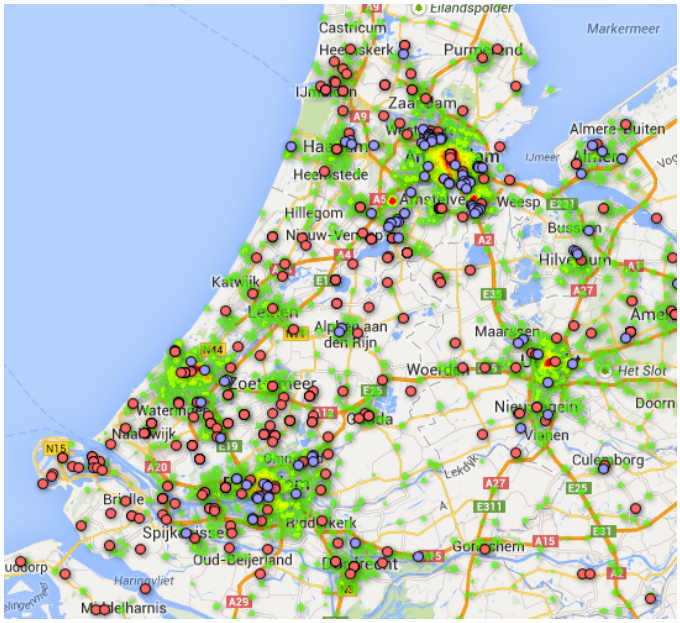


Figure 5: Three layered visualization

At the Nation level, it was possible to identify a pattern in three major cities based on the visual agreggation of the layers. Insights indicates the cities of Amsterdam, Den Haag and Rotterdam as holding a high number of communication antennas, power plants and data centers within a close range area. In specific, Amsterdam visibly shows a higher collection of ICT events.

5.2 City Level analysis

A new hypothesis is created in attempt to obtain a finer granularity on the map. This hypothesis questions if there is an agreggation of Critical Infrastructure sectors within a close range area in Amsterdam. Visualizations now take place in the city level of Amsterdam by using additional interactive layers to obtain further insights.

One of the layers contains data which provides energy measurements by neighborhoods. For a better view of data entries, shape files retrieved from the *"CBS"* where parsed to Amsterdam boundaries to create a choropleth[7]

[7]http://bl.ocks.org/mbostock/4060606

visualization. A choropleth is a style of visualization using a thematic map to shade or highlight areas in proportion to its entry values [9]. In this case, Amsterdam districts are colored based on measurements values.

Figure 6 shows an overlap of electric consumption(KWh) and CO2 emission(m3) of the city of Amsterdam during the year of 2011. This data represents the average electric consumption and average CO2 emission by companies within that area. The visualization also keeps active the layers of Power Plants and Data Center locations. The result is a four layered visualization.

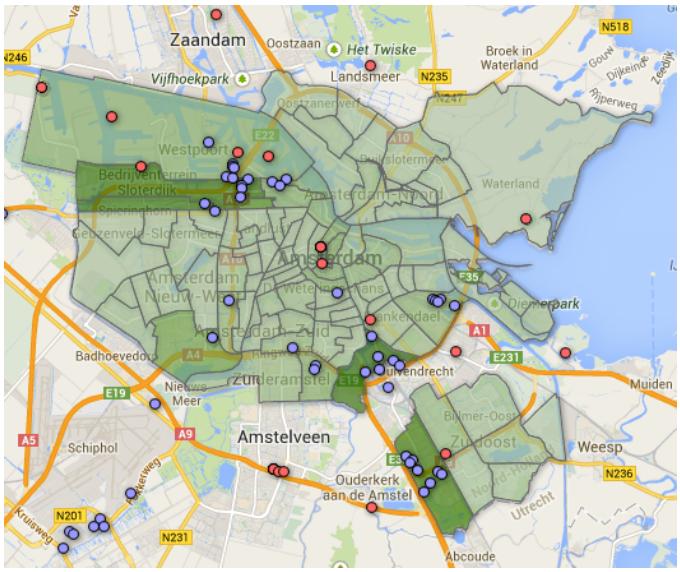


Figure 6: Four layer visualization

Apart from the limitation to provide complete colored areas (due to the nature of the available open data for geometries), it is possible to identify the high resource demanding sectors of the city by observing the scale of green. The darker green areas are also the sectors where the highest number of data centers (blue dots) and power plants (red dots) are concentrated in Amsterdam. Such areas are identified as:

- Westelijk Havengebied;
- Bedrijventerrein Sloterdijk;
- Spieringhorn;

- De Omval;
- Amstel III en Bullewijk;

Table 2 shows the relative average electricity consumption during 2011 for some of the postcodes within theses areas.

Area	PostCode	Average Business Consumption KWh (2011)
Westpoort	1041	50.411.974
	1042	47.194.887
	1045	62.935.658
De Omval	1096	39.561.140
Sloterdijk	1043	18.549.700
Amstel III/Bullewijk	1101	31.231.213
	1105	92.109.716

Table 2: Top postcode areas

An additional layer is activated to illustrate non-residential building geometries on the map. Each small dark point represents a single business related structure.

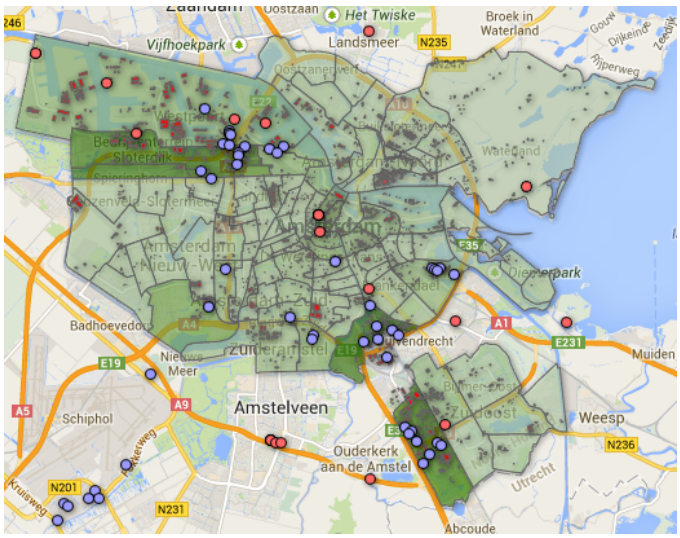


Figure 7: Four layer pattern

Figure 7 shows a pattern between datacenters, high electric consumption, power sources and business offices. Such illustration suggests the main electrical source, or backups, serving those areas. Table 3 lists the name of power

plants and servers that may have a potential relationship in energy resource handling.

Region	Power Plant	Datacenters
Westpoort	Hemweg Coenhavenweg Avi Amsterdam	EwekaDC, Tele2(Amsterdam-2), WiseXIS, ION IP, Interxion, Telecity AMS3, KPN Getronics/Pink Roccade, WeDare DataBarn1, UPC, Chello, Enertel, Sorbie, Coolwave, BT Hempoint, Gyrocenter DC2, DCA Amsterdam, Atos Origin, Equant.
De Omval	Miranda Pool	Rembrand Toren, Liander, Telecity AMS4, Verizon AMS1, euNetworks, Global Crossing, Colt.
Science Park	Diemen CSM	Equinix AM3, Plant (Matrix-1), Telecity AMS1, Nikhef.
Amstel III en Bullewijk	Hakfort Oudekerk	Verion AMS2, Switch (Internet Unie), Telecity AMS2, Equinix Virtu, Pink Roccade, Tele2(Amsterdam-1), Level3, Telecity AMS5.

Table 3: Power Plant and Datacenters

A final visualization in Figure 8 is created to address the second research question of this paper. The map view is now centralized in the Westpoort area.

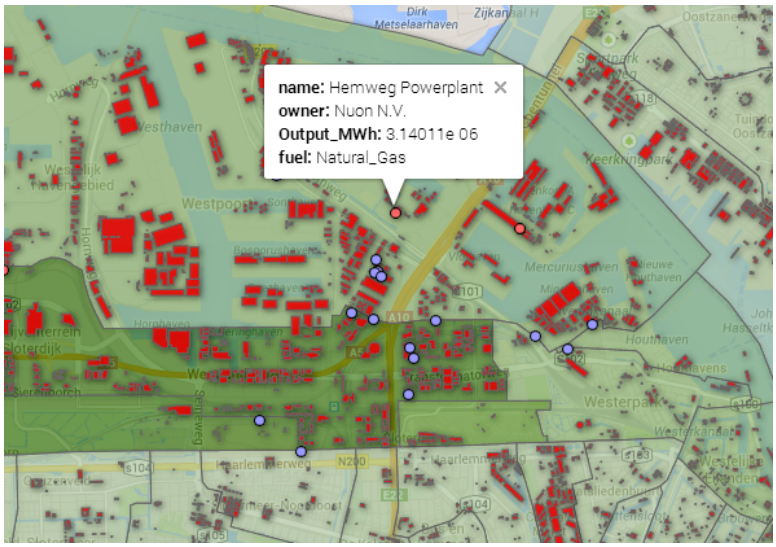


Figure 8: Hemweg Powerplant - Critical Infrastructure

From the combination of four dataset layers, the visualization shows a close range area where many datacenters and crucial source of electricity can be found. The area is also interpreted as the one consuming the most energy resources for business purposes as indicated by the green scale of the choropleth graph.
Another filter layer, containing geometry shapes, suggests the precise structure locations hosting datacenters. This last layer shows where all non-residential buildings are located, which also overlaps with the dark green areas.

5.3 Further Analysis

During the analysis phase of energy related datasets, several fields held entry value *"Afgeschermd"* to specific postcodes. Lines with that value represent areas which statistics from energy resources were intentionally foreclosed before releasing the dataset to public access. A dataset line is said to be complete available when all its values are visible. Table 4 shows a couple of lines to illustrate the occurrence of hidden values.

For geolocalization purposes, a full postcode is necessary to pinpoint a precise location on the map using Google Maps API. In the case of Dutch

PostLetter	FullAddress	Totaal verbruik Elektra (kWh)
A	AMSTERDAM,1011A	15158479
W	AMSTERDAM,1011W	350578
...	...	...
X	AMSTERDAM,1011X	Afgeschermd

Table 4: Raw energy data view

addresses, a full postcode is a combination of four numbers plus two letters e.g. 1011XA. Therefore, an approach to expand hidden postcodes to existing ones is a matter of try and error using all possible letters as the last value. Google geocode functionality is used to fetch these expanded postcodes with *"afgeschermd"* values.

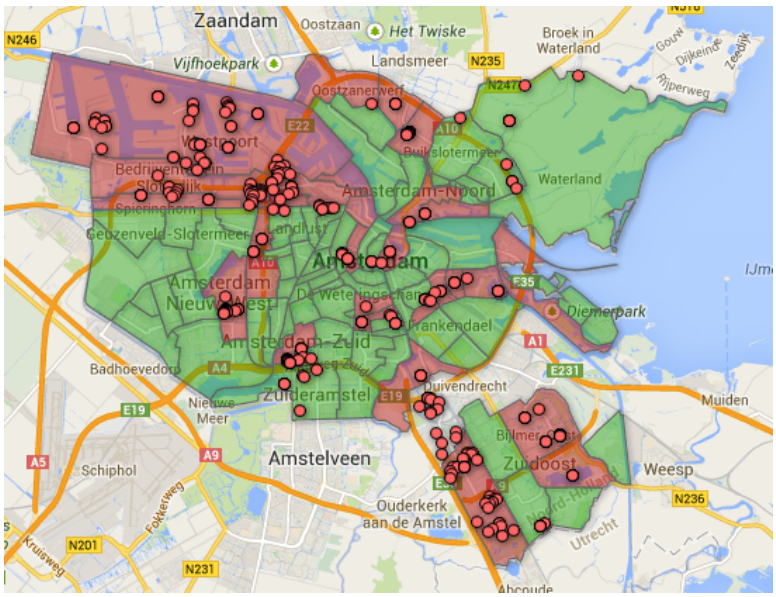


Figure 9: Amsterdam Visibiltiy - Total KWh Consumption

Figure 9 illustrates a two layered visualization that shows in red, districts were at least one postcode letter prefix was flagged with *"afgeschermd"*. All other green district areas are known to have their data open.

The red dots represent specific locations were Google geocoding succeeded to identify by using the expanded postcode approach. From this visualization it is possible to observe that the pre-identified areas with high number of datacenters and high energy consumption are also the ones with the most number of hidden entries. Thus, there is an interest to hide data availability from these areas.

6 Conclusions

This paper explained the current scenario of the Open Data initiative followed by governments and private sectors. Also, it covered in the related work, considerations regarding the nature of public data in the matter of openness versus secrecy and the issue around sensitive but unclassified information. At the core of this research, the Netherlands was subject to the proof of concept to demonstrate that it is possible to obtain relevant amount of data from public Open Data sources. By means of visual analytics techniques, it was possible to compile multiple datasets with sensitive information into a single configurable view.

One of the final visualizations provide sufficient insights to illustrate that most datacenters in Amsterdam are geographically close to the main energy sources. It also suggests which powerplants may behave as backup sources in case of service disruption. In the case of Hemweg Powerplant located in Westpoort, it is clear how critical this facility is by observing the output among in megawatts being generated and the high-resource demanding infrastructures around it.

The fact that one of the datasets contained fields with entry values flagged as *"afgeschermd"* also suggests the existing concern in not revealing sensitive information. The desire to obfuscate some areas can be seen as an institutional interest in promoting security measurements. Thus, that such information is sensitive and its disclore can be considered as a security threat.

Insights obtained by compiling public information from Open Data sources, may represent a risk to Critical Infrastructure Protection efforts. This knowledge can be obtained at any time and can be used to develop strategic plans of sabotage or even terrorism.

The matter of identifying sensitive information from Open Data database is a question of government openess versus secrecy. To some extent, such information disclosure can be something that does not depend only on government effort to maintain National Security. Most of the private organizations are included under the Critical Infrastructure umbrela by offering transportation, energy resources, supply and food chain, and water treatment. The private sector can also be held responsible in sensitive information disclosure upon users request.

Findings in this reasearch are considered not trivial to be obtained. Even within a short time frame for analysis over a specific set of data, we were able to derive interesting conclusions regarding the National Critical Infrastructures. Conclusions of this nature can be something that governments and interested parties want to avoid to be easily obtained due to National Security purposes.

7 Future Work

This research covered the scenario of Dutch Criticial Infrastructure by taking into account a specific group of datasets for detailed visual analysis.
The process described in this paper can be adapted and extended to other sources of Critical Infrastructures and can also be used to profile other nations.
Many of the visualizations provided in this research suggest other national-centric areas in The Netherlands that can be researched. For instance, Den Haag and Rotterdam can become new subjects to lower level analysis by exploring the existence of open data repositories and other public sources for those cities. Depending on the time frame available for analysis and also on the dataset reliability, these interactive visualization may produce even more detailed findings.
The existance of specific corporations profiling countries and selling such knowledge online is also known. One of these companies even advertises the quality outputs by highlighting the use of human intelligence, web crawlers, advanced news filters, social media and online databases.
This research can be extended to cover the usage of real-time sources of data. For example, URLs providing data by means of live APIs in .JSON or other formats types. These sources of data are ideal to generate analysis and visualizations on the fly due to possibility to integrate machine readable scripts. Therefore, insights can be obtained in a matter of seconds to illustrate most utilized transportation routes for example. A final visualization system can be one that automatically collects the last available and real-time datasets from reliable sources, and compile them into a single configurable interface. The output can be a multilayered map visualization offering users capabilities to rapidly acquire insights by interacting with its filters.

A wide range of software is currently available in the Big Data market to handle this urge to create visualizations from large data repositories. These software packages mostly rely on the combination of machine processing of large contents and human cognitive interpretation of data.

A full automation of this process requires the implementation of artificial intelligence and machine learning algorithms. These methodologies and techniques are mostly discussed in artificial intelligence and psychology studies [7][5].

The research in this paper confirms the possibility to derive conclusions from Critical Infrastructure regions based on the public nature of data. The second question is addressed by the implementation of a feedback loop process and a continuous visualization of data. This ongoing effort may create space to discuss in which extent this approach can be considered beneficial or dangerous. Such discussion must be left to open debate which must also consider the matter of Open Data and National Security.

Appendices

Category	Sources
Energy	senternovem.databank.nl amsterdamopendata.nl enipedia.tudelft.nl liander.nl,edsn.nl,cbs.nl
Telecommunications and ICT	antennebureau.nl nl-ix.net stat.ripe.net
Drinking water	maps.amsterdam.nl
Food	maps.amsterdam.nl
Health	maps.amsterdam.nl opendatakaart.nl
Financial sector	maps.amsterdam.nl
Surface water	maps.amsterdam.nl
Public order and safety	maps.amsterdam.nl
Legal order	maps.amsterdam.nl
Public administration	maps.amsterdam.nl
Transport	openov.nl amsterdamopendata.nl citysdk.waag.org openstreetmap.org oplaadpalen.nl gtfs.ovapi.nl trafficlink-online.nl
Chemical and Nuclear	enipedia.tudelft.nl
Air quality	lml.rivm.nl luchtmetingen.amsterdam.nl
GeoLocalization and Mapping	cbs.nl, nlextract.nl kadaster.nl, qgis.nl

Table 5: Open Data sources

References

[1] http://www.government.nl/issues/crisis-national-security-and-terrorism/protecting-critical-infrastructure.

[2] Open knowledge foundation - okfn.org.

[3] Roba Abbas. The risk of public data availability on critical infrastructure protection. In *The First Workshop on the Social Implications of National Security*, 2006.

[4] Fahimeh Alizadeh and Razvan C Oprea. Discovery and mapping of the dutch national critical ip infrastructure. 2013.

[5] John Robert Anderson, Ryszard Spencer Michalski, Ryszard Stanisław Michalski, Thomas Michael Mitchell, et al. *Machine learning: An artificial intelligence approach*, volume 2. Morgan Kaufmann, 1986.

[6] Edson Kowask Bezerra, Emilio Tissato Nakamura, and SL Ribeiro. Critical telecommunications infrastructure protection in brazil. In *Critical Infrastructure Protection, First IEEE International Workshop on*, pages 11–pp. IEEE, 2005.

[7] Mark H Bickhard and Loren Terveen. *Foundational issues in artificial intelligence and cognitive science: Impasse and solution*, volume 109. Elsevier, 1996.

[8] ESRI ESRI. Shapefile technical description. esri. *INC, http://www. esri. com*, 1998.

[9] Jeffrey Heer, Michael Bostock, and Vadim Ogievetsky. A tour through the visualization zoo. *Commun. ACM*, 53(6):59–67, 2010.

[10] Daniel A Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas, and Hartmut Ziegler. *Visual analytics: Scope and challenges*. Springer, 2008.

[11] Eric Luiijf, Albert Nieuwenhuijs, Marieke Klaver, Michel van Eeten, and Edite Cruz. Empirical findings on critical infrastructure dependencies in europe. In *Critical Information Infrastructure Security*, pages 302–310. Springer, 2009.

[12] John D Moteff and Gina M Stevens. Critical infrastructure information disclosure and homeland security. DTIC Document, 2002.

[13] Andreas Neumann and Marco Hugentobler. Open-source gis libraries. In *Encyclopedia of GIS*, pages 816–820. Springer, 2008.

[14] Deborah Nolan and Duncan Temple Lang. Keyhole markup language. In *XML and Web Technologies for Data Sciences with R*, pages 581–618. Springer, 2014.

[15] Steven M Rinaldi, James P Peerenboom, and Terrence K Kelly. Identifying, understanding, and analyzing critical infrastructure interdependencies. *Control Systems, IEEE*, 21(6):11–25, 2001.

[16] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343. IEEE, 1996.