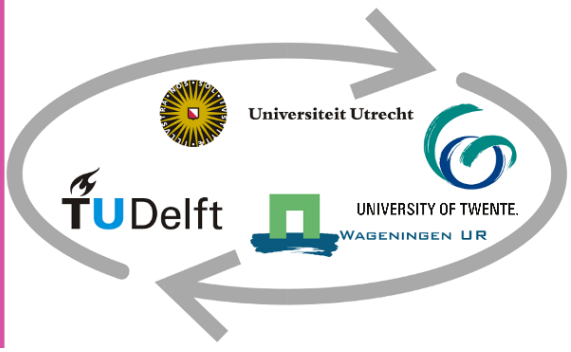
GIMA
Geographical Information Management and Applications

Towards a sustainable geoprocessing environment at Statistics Netherlands through performance benchmarking

Sandra Desabandu
June 2015

Professor: prof. dr. ir. P.J.M. van Oosterom (TU Delft)
Supervisor GIMA: ir.E.Verbree (TU Delft)
Supervisor Statistics Netherlands: ir. Pieter Bresters



Towards a sustainable geoprocessing environment at Statistics Netherlands through performance benchmarking

A research into the applicability of performance benchmarking within proprietary off the shelve GIS of Esri and lessons learned of similar organizations for decision making on a modern, sustainable geo-ict infrastructure.

Date: 01.06.2015

Author: Sandra Desabandu

Student number: 3686841 (University of Utrecht)

E-mail: cello2529@gmail.com

Supervisor GIMA: ir. E.Verbree

E-mail: E.Verbree@tudelft.nl

Supervisor Statistics Netherlands: Peter Bresters

E-mail: p.bresters@cbs.nl

Professor: Prof.dr.ir. P.J.M van Oosterom

E-mail: P.J.M.vanOosterom@tudelft.nl

Preface

Computer performance is a common problem for anyone working regularly with often accepted as a daily annoyance. For anyone working with Geographic Information Systems, performance can be a significant bottleneck that slows down analysis and visualization of geographic data. For an organization such as Statistics Netherlands, that performs complicated spatial calculations with datasets of (in many cases) 8.000.000 records, satisfactory performance is almost a business-critical requirement. This has made the set-up and conducting of this research project especially challenging because scientific research within such an important real life process of a large public agency requires to step back at first, to find your way within related research and to set up a research strategy. At the same time, it is equally important to use the knowledge of the very motivated, experienced and skilled staff members of Statistics Netherlands. In this environment, a performance benchmark had to be devised, within the constraints of a real life organization and its results had to lead to an advice on the future geo-ict infrastructure. Therefore, the research had to be performed from different angles: technical performance factors to be used for the benchmark, but also organizational factors that could influence the advice. Ex The real data scenarios will be executed on the big data computer in ArcGIS 10.2. This is not directly comparable with ArcGIS 10.1.

periences from other organizations have therefore been an important part of the thesis.

I would like to thank Statistics Netherlands for providing the opportunity to conduct my thesis research and Pieter Bresters as my supervisor from Statistics Netherlands for valuable support and advice and introducing me very well into the Spatial Statistics department. I'm very grateful also for all the very useful input and ideas from the staff members of Spatial Statistics and head of the department Eric Fokke for taking care so well of the organization of the internship.

GIMA has been a great place to learn about the vast field of Geographic Information and about scientific research. I owe many thanks to my supervisor Edward Verbree and my supervising professor Peter van Oosterom for their constructive feedback.

Also, I would like to acknowledge the role of Esri: Without support, input and ArcGIS Pro testing facilities from Esri, this research would not have been possible. I want to thank especially Ernst Eijkelenboom for all support. The draft version of the thesis has been provided to Esri for reviewing.

Very valuable information on geoprocessing performance experiences have been shared by a number of organizations where geoprocessing of large datasets is an important process and ESRI software is used:

- PBL (Martijn Spoon and Arjan van der Put)
- Statistics Portugal (Bartolomeus Schoenmakers)
- Statistics Italy (Rossella Molinaro)
- United States Geological Survey (Curtis Price)

Additionally, I also would like to acknowledge the contribution of Aris (Eddy Scheper and Anke Keuren) as an organizations that provided insight into migration from ArcInfo Workstation to ArcGIS Desktop.

Conducting such a research is interesting and fun, but also very heavy, especially if it is in combination with the "regular" work and family. Without support of my family, especially my husband, children, parents and parents in law, this task would have been impossible. My husband Marijn Zuurmond stands out especially: foremost for his patience and loving support.

Sandra Desabandu

Abstract

The research project has aimed to support decision making regarding possible transition towards an up-to-date, sustainable geo-ict architecture based on an analysis of technical factors that influence performance of geoprocessing tools at the spatial team of Statistics Netherlands and experiences at comparable organizations. The current geoprocessing environment is ArcInfo Workstation and ArcGIS Desktop 10.1, although most production processes are conducted in ArcInfo Workstation and its scripting language AML. Support of ArcInfo Workstation will not be continued after 2015. Therefore, migration of these processes to ArcGIS Desktop 10.1 or higher version of this product suite has to be considered. Consequently, the following research question has been formulated:

Which alternatives to the current geo-ict infrastructure can be proposed for Statistics Netherlands that meet performance requirements of its geoprocessing activities and are suitable for implementation within the organizational constraints of Statistics Netherlands?

Based on studying the workload, resources and performance bottlenecks at Statistics Netherlands, a benchmark has been developed evaluating four geoprocessing tools (UNION, INTERSECT, DISSOLVE and NEAR) to be tested on scalability (using synthetic data), impact of optimization factors available in ArcGIS Desktop 10.1 and impact of big data hardware and new software releases ArcGIS 10.2 and ArcGIS Pro (using real data). A number of administrative processing tools (SUMMARY STATISTICS, FREQUENCY, CALCULATE and JOIN FIELD) has been additionally included for tests with the big data hardware and the new software releases. The results of the benchmark have been combined with migration experiences of ArcInfo to ArcGIS Desktop or ArcSDE from other organizations and held against possible internal and external restrictions and trends. These organizations are Statistics Portugal, Statistics Italy, United States Geological Survey and PBL (Netherlands Environmental Assessment Agency).

The benchmark results showed different results per tool: Whereas UNION and INTERSECT show the same performance in ArcGIS desktop 10.1 as in ArcInfo Workstation, the DISSOLVE and the JOIN are considerable slower in all higher desktop versions. The selected geoprocessing tools UNION, INTERSECT, DISSOLVE and NEAR did not show improvement with the use of available optimization factors (spatial index, spatial sort, parallel processing environment, compacting and compressing) in ArcGIS 10.1, although the use of optimization factors has not been exhaustive. The most remarkable results showed a decline in performance, for example compression of the input datasets. The NEAR shows no difference between ArcInfo Workstation and ArcGIS 10.1 on the fat client, but showed a big improvement in ArcGIS 10.2 and ArcGIS Pro on the big data computer. The improvement of the NEAR implicates that the algorithm of the tool has been redesigned. The other tools, also a number of administrative processing tools tested for Statistics Netherlands, did not show substantial improvement within a big data hardware environment and within ArcGIS Pro. The lessons learned that have been obtained by the interview and questionnaires can be formulated in several points:

1. The transition has been a gradual process for the respondent organizations, but for Statistics Netherlands the urgency to migrate is higher because of the high number of production processes written in AML that need to be migrated because of the production load in AML and the product cycle of ArcInfo Workstation.
2. The influence on ICT infrastructure decision making in terms of facilities for geoprocessing will depend on the user base and the organization of the GIS users:
3. The PBL shows interesting options in optimization, but also has a more flexible (Geo) ICT infrastructure.

4. Open Source or other proprietary software is largely new territory and not really investigated on functionality and performance.
5. For most the Statistical agencies and the PBL some part of the migration has been outsourced, at least the initial phase of the migration.
6. The evaluation of the migration to a higher ArcGIS version is mixed and varies per tool or model. It is also dependent on the user needs of the organization.

To provide a migration advice that also takes into account future developments, a number of trends in geoprocessing technology have been described, such as cloud computing, MapReduce/Hadoop. Applying these technologies would require a substantial investment in knowledge and adaption of production processes. Moreover, solutions that affect the ICT infrastructure as a whole, like cloud computing or could affect the results of production of statistics by using algorithms like MapReduce are less likely to be implemented within the near future. Based on the previous information, the following conclusion can be stated: Given the limited scalability and end of support of ArcInfo Workstation, migration to a more up-to-date infrastructure is unavoidable. Three alternatives are possible:

- Alternative 1: Keep small scale infrastructure within ArcGIS Desktop suite an future ArcGIS Pro Spatial Statistics stays within the ArcGIS Desktop suite but should consider quicker migration to 10.2 or rather 10.3 because more performance problems are resolved within these versions (according to ESRI) and because the possibilities to optimize ArcGIS Desktop 10. 1 remain very limited. This will mean that Spatial Statistics will need less expertise in-house on performance optimization, but will remain dependent on ESRI product development processes. This dependency could be countered by cooperating more closely with other ArcGIS users nationally and internationally to press for the needed improvements.
- Alternative 2: Invest in a sustainable (geo) ict environment with other departments
  To be more in control of performance optimization means that a more configurable Geo ICT infrastructure is needed. Spatial Statistics is too small to make such an investment in infrastructure and needed expertise (“humanware”). Therefore it should cooperate with other departments at SN who have a need for spatial data analysis. Such a (geo) ict environment could include several components: a spatial database such as Oracle combined with ArcSDE, an open source spatial database such as PostGIS or a file-based environment. Additionally, extension of the benchmark towards spatial databases would be needed. This need is especially apparent if the volume of data is expected to grow in future.
- Alternative3: Organize innovative projects with internal and external partners as a complementary step: In addition to the more short-term solution Spatial Statistics should start a number of projects dedicated to the application of new geo-ict technologies. For example, a project involving Spatial Statistics and other teams in cooperation with the innovation laboratory.

The conclusions have also resulted in a number of recommendations that are aimed at improving the benchmark and at helping the migration process of Statistics Netherlands.

Acronyms

AML
Arc Macro Language, scripting language of ArcInfo Workstation

ArcGIS Desktop

ArcSDE
Spatial Database Engine used as a component of ArcGIS Server to manage data in a spatial database.

BAG
Basisregistratie Adressen en Gebouwen (Key Register Addresses and Buildings)

BRT
Basisregistratie Topografie (Key Register Topography)

CIFS
Common Internet File System (CIFS) is a protocol that programs make requests for files and services on remote computers on the Internet. CIFS uses the client/server programming model. A client program makes a request of a server program (usually in another computer) for access to a file or to pass a message to a program that runs in the server computer. The server takes the requested action and returns a response (Rouse, Margaret, 2005).

CPU
Central Processing Unit

CUDA
Compute Unified Device Architecture

GPU
Graphics Processing Unit

LAN
Local Area Network

OGC
Open Geospatial Consortium: organization that defines open standards for geographic data and geographic applications

SN
Statistics Netherlands

SCCM
System Centre Configuration Manager

VBA
Visual Basic for Applications

WAN
Wide Area Network

List of figures

List of tables

CONTENT

Chapter 1: Introduction

This chapter will describe the research motivation in section 1.1 and forthcoming research objective (described in section 1.2). Out of the research objective, main research question and sub questions will be deducted in section 1.3. Section 1.4 will provide an overview of expected results, whereas section 1.5 will form the scope of the research.

1.1 Motivation

Spatial calculations are often very time consuming because of the complexity of spatial data models and large size of datasets. Some examples are overlay calculations of a land use polygon dataset with a road dataset or a DISSOLVE of a buildings dataset. Therefore, performance of these calculations is a critical success factor for geoprocessing of large spatial datasets. For organizations that deal with large (spatial) datasets, performance plays an important role in decision making regarding the (geo-) information infrastructure. It is important to base such a decision on valid information regarding the performance of certain architecture scenario's, because systems migration is a complicated, long process with high chance of failure if the new architecture does not answer business critical needs of an organization. This is also the reason for many organizations to hold on to older legacy systems that are not supported and developed further but that still provide mission critical services (Al-Azzoni et al. 2011).

Governmental agencies are important organizations that create process and store heavy datasets. However, with the growth of spatial and spatio- temporal data, different data formats (for example Lidar) and higher use and production of spatial data in everyday life, more organizations that process spatial data regard performance as an important or even business critical aspect of their (geo) ICT infrastructure. SN is a so-called ZBO (Zelfstandig Bestuursorgaan), a governmental service that operates without hierarchical relationship to a minister. However, the minister of Economic Affairs is responsible for the financial means of SN. An independent commission, the Central Commission for Statistics Netherlands, defines the statistical research agenda for the following period an monitors quality and reliability of the research activities (Statistics Netherlands, 2015a).

The main task of Statistics Netherlands (abbreviation SN) is the production and publication of reliable, coherent statistical information, addressing the information needs of Dutch society. Additionally, it is also tasked with cooperation on European statistical data. To be able to conduct statistical research, many data have to be collected about citizens, companies and organizations. Part of these (micro-) data is very privacy sensitive. A special Statistics Netherlands act has been installed that provides strict regulations on how to deal with privacy sensitive (Statistics Netherlands, 2015b) data.

Approximately 2000 staff members are working at Statistics Netherlands, divided between two locations: The Hague and Heerlen. The organization is hierarchically subdivided into divisions, sectors and teams. Figure 1 shows an organization schema of Statistics Netherlands. The division socio-economic and Spatial Statistics (marked in red font) incorporates 7 sectors of which the sector Environmental, Energy and Spatial Statistics houses a number of teams: Environment, Energy, Real Estate and Spatial Statistics.

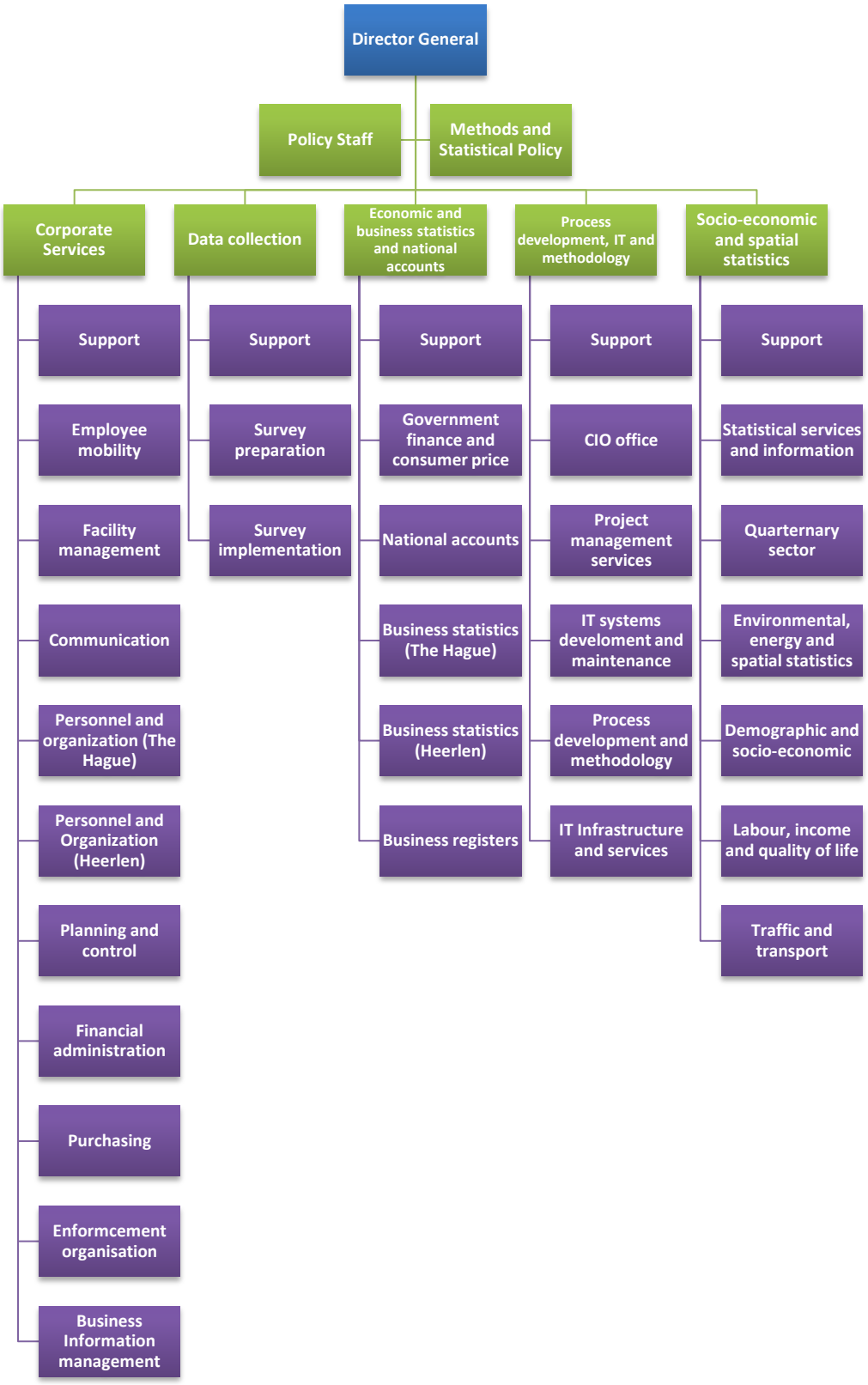


Figure 1: Organization Chart SN (SN, 2014)

Spatial Statistics is a relatively small team, compared to the scale of the organization. Four staff members have the role of project manager for the following production processes: the standard products, data management, web services, EURO Stat projects (e.g. INSPIRE) and custom products. Part of the statistical researchers has a GIS background, conducting GIS analysis, programming, while other researchers are tasked with European policy on boundaries and analysis of temporal changes of addresses within the Netherlands. Three staff members are working as operators, their main task is data entry and imagery interpretation.

The department "Regio and Ruimte" (Spatial Statistics) produces its own statistics, but also supports other (light) users throughout the organization with the use of GIS tooling. These users are especially present in the same sector as Spatial Statistics (Environmental, Energy and Spatial Statistics), but also in sectors Labor, Income and Quality of life Statistics, Traffic and Transport Statistics, Quaternary Sector Statistics, Statistical Services and Information and Business Statistics (Goedhuys, 2014). More on the use of the diverse GIS applications will follow in section 3.1. Spatial Statistics is responsible for the production of spatial statistical products. Together with the regular products, a number of custom products are delivered. The most important regular products of Spatial Statistics are (see Appendix 1 for detailed description):

- Bestand Bodemgebruik (BBG – Land use map for the Netherlands)
- Financiële Verhoudingswet (analysis of spatial features per municipality, used as a basis for amount of subsidy provided by the Ministry of Home Affairs)
- Nabijheidstatistieken (Proximity Statistics)
- Wijk- en Buurtkaart (Municipality- District- and Neighbourhood statistics)
- Vierkantstatistieken (Grid-based statistics)
- Bevolkingskernen (Urban agglomeration)
- Geoservices

These products are widely used by internal and external customers: Magazine Elsevier uses the proximity statistics regarding public transport, healthcare and recreation facilities in combination with other data to determine the "best municipality" of the year (Elsevier, 2012). This information is also used by real estate agents. The "Financiële Verhoudingswet" calculations are crucial for subsidies from the national government to municipalities. A number of products are available via PDOK[1], a portal that disseminates public geodata via WMS or WFS: Land Use, Grid Statistics, Urban Agglomeration and Municipality-/District-/Neighbourhood Statistics. These products, the widespread use, and the legal background of the "Financiële Verhoudingswet" mean that there is a high quality standard that has to be maintained as well as timeliness of the results.

Most of the products are the result of very time consuming analysis and processing of large spatial (micro) datasets. Among the datasets that are mostly used are the (spatial) Dutch key register datasets such as the BAG (key register of addresses and buildings), the GBA (key register of residents), the BRT (topographical key register) or the NHR (key register of the chamber of commerce), complemented by imagery of the Netherlands and location data of several sectors within the Netherlands. An administrative join between the personal records database (GBA) and the key register of buildings and addresses (BAG) in combination with spatial overlays between layouts of territories often takes more than 24 hours on the standard fat clients at Statistics Netherlands.

Most products are based on the use of different geoprocessing tools in Arc Workstation, written in Arc Macro Language (AML). ArcInfo uses old file formats like Coverage, Info and shapefile. ArcGIS Desktop 10.1 is also part of the production process, but the staff members of Spatial Statistics have experienced better performance of geo processing tools in Workstation and AML, although no further research has been conducted on the reasons and on possible methods to improve performance in ArcGIS Desktop 10.1. Migration to ArcGIS Desktop might be necessary for various reasons:

[1] Public Services on the map

- Knowledge of AML and Workstation is becoming very scarce, therefore maintenance of the tools will become problematic.
- Support of the software producer ESRI for Workstation will stop after the introduction of Windows 8 and ultimately after 31.12 2015 (ESRI, 2014a).
- ArcInfo Workstation also shows limitations in loading current key register datasets.

These reasons show that the current situation may provide better performance of geoprocessing, but may not be sustainable for long term future production processes. Therefore, the developers within the spatial team have already started to redesign a part of the production processes for ArcGIS 10.1 (the process for the land use dataset).

The redesigned process used a different method is used to detect changes in land use and to edit the data. For this reason, it is not possible to compare performance of the complete process between the old and new situation: other geoprocessing steps and input data are used in both methods. It is, however, possible to examine performance on the level of geoprocessing tool, using different factors that contribute to performance and a selection of input datasets in different population sizes. The spatial team has already identified a number of viable factors that have influence on performance, for example data format, population length and width (number of records and number of attributes), time/day of tool execution and software. More factors will be described in detail in chapter 4. The number of performance factors is already high and more important factors are expected to be added as a result of literature research and interviews. Therefore, prioritizing the factors will be equally important.

To be able to take a well-founded decision about possible migration from Workstation to an up-to-date, sustainable, geo-ict architecture, for example ArcGIS Desktop 10.1 or higher without AML, Statistics Netherlands wants to assess if such a migration will affect performance of analysis processes and which measures would be advisable to keep an acceptable level of performance. This implies performance evaluation of geoprocessing steps in combination with different performance factors/variables, but also using knowledge and best practices of other organizations that have dealt with migration from a legacy spatial DBMS like ArcInfo Workstation towards a more up-to-date and sustainable geo-ict-architecture. Additionally, Statistics Netherlands has to look for a viable solution if the new ArcGIS version is not fulfilling the performance needs of the CBS. The software used for geoprocessing is, of course, an important factor that can influence performance. Software solutions for big data like Hadoop or statistics like R have developed a spatial extension and have been tested in diverse studies, for example in Aji et al. (2013). While the results of the performance evaluation may point towards a certain solution for Statistics Netherlands, the “real life” setting of the research at Statistics Netherlands also calls for implementability of such a solution. Therefore, the organizational constraints have to be considered.

1.2 Research objective

The aim of this research project is to support decision making regarding possible transition towards an up-to-date, sustainable geo-ict architecture based on an analysis of technical factors that influence performance of geoprocessing tools at the spatial team of Statistics Netherlands and experiences at comparable organizations. Comparable organizations are governmental agencies that process large spatial datasets, such as, e.g. the PBL (Planbureau voor the Leefomgeving), statistical organizations in other countries, semi-public organizations such as TNO, but also commercial organizations that make value-added products based on large datasets. The results of the research should be evaluated against the backdrop of organizational constraints (for example business critical processes, organization of ICT), commercial decision making of software vendors and especially the level of performance that is acceptable for Statistics Netherlands and lead to a viable scenario for the geo-ict architecture of the organization.

1.3 Research questions and expected results

Based on motivation and research objective, the central research question has been formulated as follows:

Which alternatives to the current geo-ict infrastructure can be proposed for Statistics Netherlands that meet performance requirements of its geoprocessing activities and are suitable for implementation within the organizational constraints of Statistics Netherlands?

To be able to answer the central question, the following sub questions have to be answered first:

1. Which technical as well as organizational performance bottlenecks can be identified during geo-information processes at Statistics Netherlands and how are they currently dealt with?
2. Which (technical) performance factors can be identified for geoprocessing based on literature examples and on experiences at Statistics Netherlands (a number of factors have been identified by SN) and how can these factors be prioritized and evaluated with the design of a benchmark, based on the performance requirements of Spatial Statistics?
3. What are the results and conclusions of the performance evaluation tests?
4. What are geo-ict related trends (e.g role of performance in following releases of ArcGIS or other proprietary or open source software) that could influence the recommended migration scenario for Statistics Netherlands?
5. What are lessons learned from organizations with a profile similar to Statistics Netherlands that process large spatial datasets and can provide information on performance evaluation and its role within decision making related to geo-ict infrastructure?
6. Which (geo) ICT scenario is most suitable for the organization described in the case study including implementation steps and quick-wins?

These research sub questions can be translated into the schema shown in Figure 2 where the violet block "recommendations" is the final result, based on the other blocks:

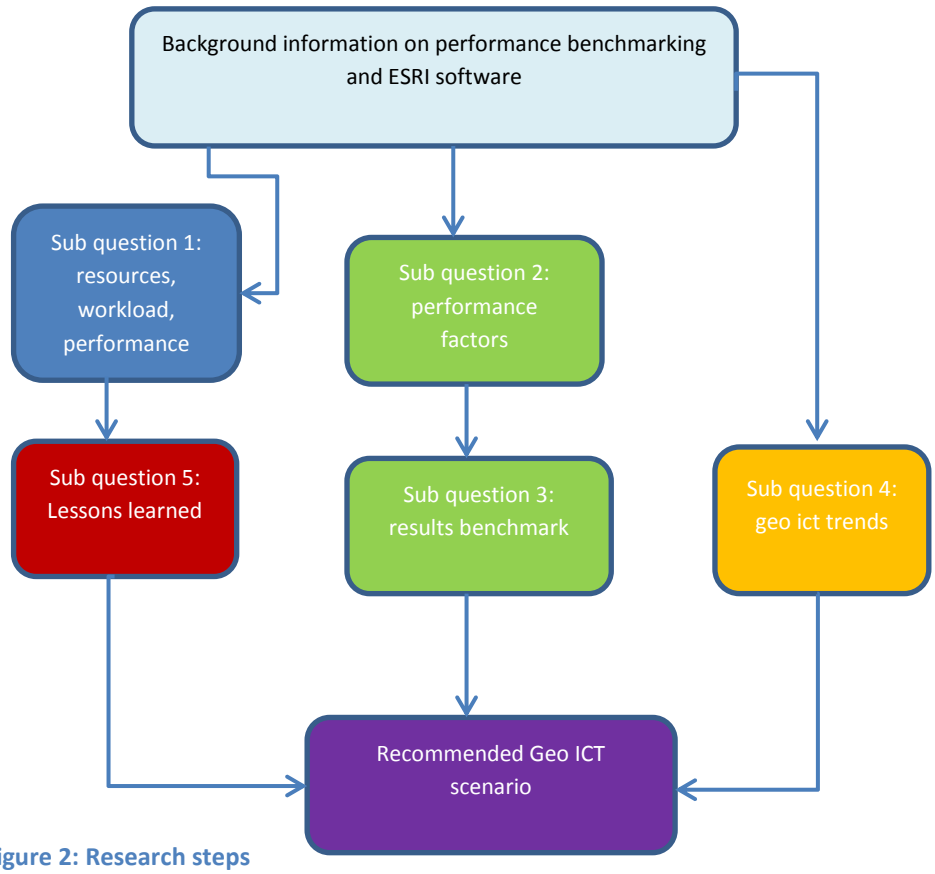


Figure 2: Research steps

- Blue: Sub question 1 (Chapter 3):
  A global description of Statistics Netherlands as an organization will be provided, followed by the position of Spatial Statistics as well as its products, application landscape, workload and resources. Based on this information, organizational as well as technical performance bottlenecks can be analysed.
- Green: Sub questions 2 and 3 (Chapter 2, 4, 5 and 6):
  A performance evaluation method/benchmark applicable for measuring the influence of certain technical factors on a selection of geoprocessing tools and the testing results, visualized in performance charts per performance factor and geoprocessing tool and interpreted in relation to the testing process
- Orange: Sub questions 4 and 5 (Chapter 8 for geo-ict trends)
- Red: Sub question 6 (Chapter 7): An analysis of lessons learned/best practices of organizations with a profile similar to Statistics Netherlands
- Violet: Sub question 7 (Chapter 9 and 10): Conclusions and Recommendations for a geo-ict scenario, based on research results, including quick wins for the spatial team of Statistics Netherlands

1.4 Scope

The benchmark will focus on the technical factors (for example: software, indexing, data size, parallel processing and hardware) that can affect performance of geoprocessing tools. The organizational bottlenecks should be used to create a realistic migration scenario for Statistics Netherlands.

A number of performance factors have already been stated by Statistics Netherlands and even more factors could be derived from scientific literature. Using too many factors could introduce bias to the testing results, therefore prioritizing factors and weighing them in importance will be vital to achieve clear results and to finalize the research within the provided time. The factors that can be tested at Statistics Netherlands within ArcInfo Workstation and ArcGIS 10.1 also have priority.

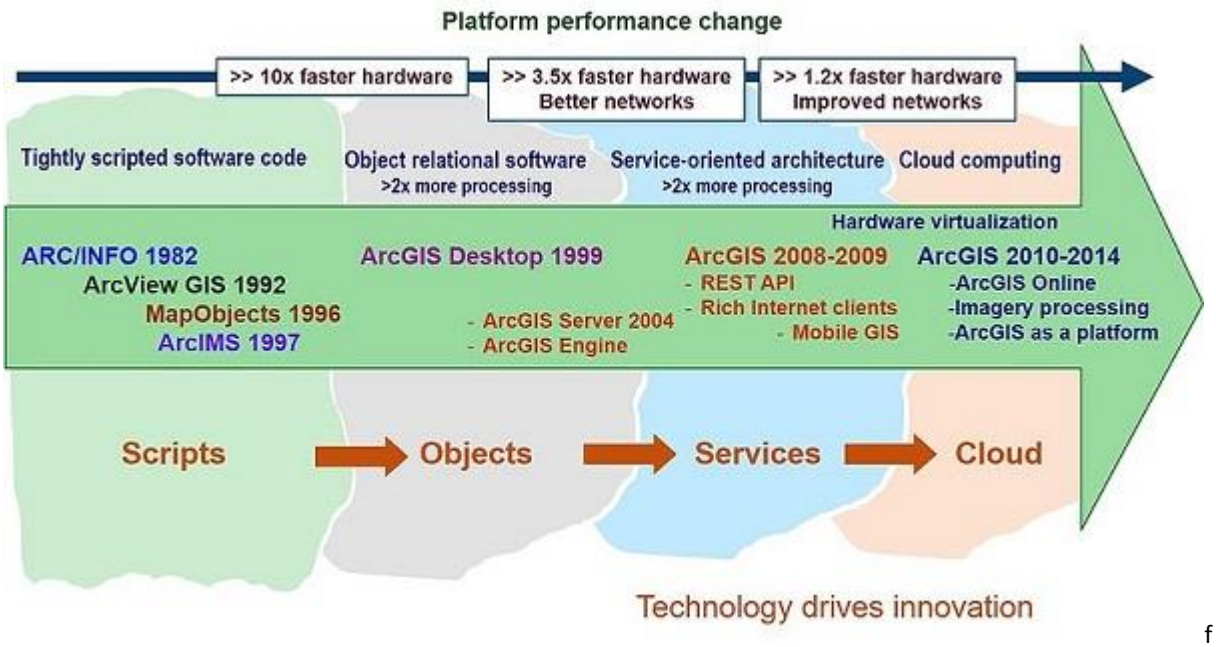
The benchmark will not include ArcGIS in combination with spatial databases such as Oracle or PostgreSQL nor will it cover other proprietary or open source GIS software. It will also limit itself to the data formats coverage and file geodatabase.

Chapter 2: Background Information of ESRI software and performance benchmark theory

To understand the implications of software migration from ArcInfo Workstation to ArcGIS Desktop, a clear picture of the involved applications and their development history is important. This will be provided in section 2.1. Section 2.2 will deal with the methodology of benchmark development, assessment of workload and resources. Developing a benchmark calls for a methodological foundation on the meaning of performance, performance metrics, benchmark development, workload and resources analysis. This will be covered in sections 2.3 through 2.8.

2.1 ESRI software development

To place the applications that are used at Statistics Netherlands within its historical context, a short time line of the product development of ESRI will be described. Figure 3 shows the development of the ESRI software from the early 1980's through 2014 from a script based via object based towards service and cloud based environment. Growth in hardware performance has stimulated this development. The development in performance has been mainly dependent on faster hardware during the transition from scripts towards objects. Higher network performance emerged from the object towards a service oriented environment and played an important role from a service-oriented architecture towards cloud computing.



f

Figure 3:Evolution highlights of ArcGIS software from 1982 to present (Peters, 2014)

The development of ArcInfo has been based on the "toolkit principle", meaning that geographic features are defined as objects, with geoprocessing operators to perform GIS analysis on these objects. This is in contrast to the spatial database approach (Morehouse, 1989). The application has been designed in modules for data structure, device interface, processing and program. Each module has been designed independently by one programmer to make the application more flexible and expandable. ArcInfo Workstation has a command-line interface and its tools are based on AML.

The data model has been developed to be compliant with the geoprocessing tools. The most important part of the data model for ArcInfo is the coverage (more information provided in section 2.2). The programming/scripting language of Workstation is AML, a runtime interpreted language. The last version (version 10) of ArcInfo Workstation has been issued together with ArcGIS Desktop 10.0. There will be no further development although Workstation can be used in combination with higher desktop versions. The application will retire as of 31.12 2015 (ESRI, 2014a).

Whereas ArcInfo Workstation has been developed for GIS professionals, the need to develop a desktop, suitable for users that are not geo-information specialists, emerged. ArcView has been the first step, developed in the beginning of the 1990's, first as a more user friendly graphical interface, compatible with Windows, to view geographic data in contrast to the command line driven ArcInfo. Gradually, more functionality has been added to enable users to conduct more complex spatial analysis. For several reasons, ArcView still has a number of dedicated users for reasons of cost (low licence costs), legacy files and scripts as well as functionality: Some editing and overlay functionality seems to be easier and quicker than in ArcGIS Desktop (Morais, 2008).

ArcInfo and ArcView work with file based datasets, but the need to manage and share spatial data lead to the development of ArcStorm and later of ArcSDE, a middleware component that is used in combination with spatial database of Oracle, Informix, IBM DB2, SQL Server and Sybase (Guan, 2006). The need to share geo-information without multiple copying and storage of data also lead to the development of an environment that facilitated web services. This led to the development of ArcIMS around the time of the release of ArcSDE (Peters, 2008).

The software development also changed in the early 1990's, from traditional scripting to object component coding (Peters, 2008). The chosen language is the Microsoft Common Object Model (COM). In 1999, ArcGIS desktop 8, developed in COM, was issued (Guan, 2006). It was the start of a new desktop product, combining ArcView interface with the heavy functionality of ArcInfo Workstation. With ArcGIS desktop 9.3, a number of new components have been introduced such as ArcCatalog, an application to manage GIS files and the toolbox containing a collection of different data management or geoprocessing tools. ArcGIS 10 marks another milestone, the integration of ArcCatalog and ArcMap and the transition from VBA[2] to Python as a scripting language and VB.NET or C# as a developing language (ESRI, unknown). Around that time, in 2008, service-oriented architecture became more important. It enabled maintenance of data integrity, reduction of storage costs, as well as the use of map services on a mobile device. New releases are often providing new functionality or bug fixes, but also the redesign of tools to improve performance: For the release of 10.2, for example, the GENERATE NEAR TABLE and the NEAR tools have been rewritten: According to ESRI, they have been rewritten to be "dramatically faster " (ESRI, 2014d). New versions of the software or new service packages for a certain version are used to address bugs and performance problems and to introduce new functionality. In 10.3, also a number of performance related improvements have been made, compared to 10.2.2 (ESRI, 2015).

The current decade is marked by the development towards cloud computing and the use of ArcGIS as a platform, not only for use as desktop software, but also as a platform to exchange data. The new desktop GIS ArcGISPro will be installed and updated via the platform, which means the end of official versions and packages. This application is the most recent desktop GIS product of ESRI, aimed at professional use. Since January 2015, version 1.0 has been released. It is completely redesigned on different levels: interface, workflow, use of 2D and 3D, scripting and use of hardware resources. The user interface is very different from ArcGIS Desktop: it uses a "ribbon interface" that activate the menu tabs or buttons that are needed within a certain process. A project is used in ArcGISPro to keep resources such as maps, layouts, connections to databases at the same place and to enable cooperation on these resources between different project team members (ESRI, 2015).

Python 3.4 is used in ArcGIS Pro to automate geoprocessing steps. It is important to consider that ArcGIS scripting is still authoring Python. The current Python version used with ArcGIS 10.1 is Python 2.7. The geoprocessing tools can be executed on the fly, via the toolbox as a model or a script tool, which can contain a python script, a batch or executable file. ArcGIS Pro does not support coverages and INFO files, nor does it

[2] Microsoft Visual Basic for Applications

support conversion to and from coverage. ArcGIS Desktop 10.3 is still supporting this possibility. Performance improvement with ArcGIS Pro has been indicated by ESRI because of (Desabandu & Eijkelenboom, 2014):

- GPU drive graphic interface, therefore the application is visualizing spatial data much faster
- statistical (raster) calculations that have been redesigned (there is no complete list with redesigned tools available). It is not known whether vector tools have been redesigned.
- more efficient memory use as a 64 bit application

It is possible to run ArcInfo Workstation 10, ArcGIS Desktop 10.1 en ArcGIS Pro concurrently on one system. The ArcGIS Pro release will take place with the release of ArcGIS Desktop 10.3. The new development environment that can be used to extend ArcGISPro is Net SDK[3], following up ArcObjects. The SDK does not support custom geoprocessing functions (Elkins Jr & Macleod, 2014).

2.2 ESRI data structures

2.2.1 Coverages

Introduced by ESRI as an innovative data format with topology in the 1980's to create a GIS format containing topology in contrast to AutoCad files (Van Dyke, 2009). A coverage dataset is stored in the computer as a directory (workspace) with the coverage file name. The directory contains different feature classes. A number of them is shown in Figure 4, whereas the meaning of those different types of feature classes is explained in Table 1.

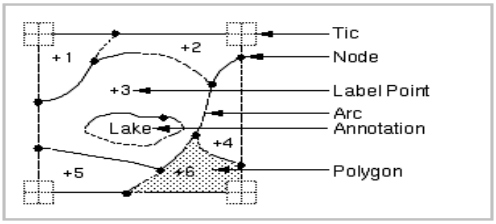


Figure 4: Coverage data structure (ESRI, 2010c)

TIC	Bounding box poins of feature class extend
Node	intersection point where two or more arcs meet
Label	Label points, stored as a separate feature class
Arc	a straight line vector that start and end at a node
Polygon	Polygon feature class

Table 1: Coverage feature classes

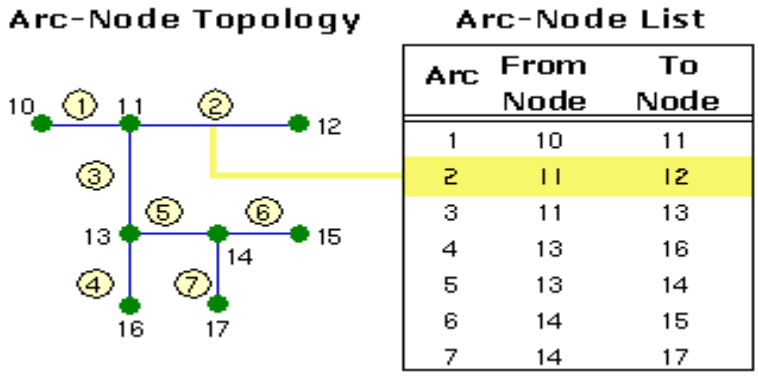
Next to the feature classes, the attribute data are stored separately in tables (INFO). Viewing the feature class in ArcCatalog, the attribute data seem to reside in the same table. Specific feature attributes are stored in .adf files. Coverages are based on explicitly stored topology which means that information on the topological relationship is stored in tables that are related to the represented features (Batcheller, 2007). The topology can be enforced and updated with the ArcInfo commands CLEAN and BUILD (Theobald, 2001). Explicit topology eliminates the necessity to compute the topology on the fly and could therefore reduce computing capacity.

There are two main topological principles that are stored within the coverage data structure:

[3] Software Developer Toolkit

a) **Linear network topology**, called “arc-node topology principle” by ESRI, which is the basis for network calculation

This principle stores the connectivity between the line features, which are called arcs within a topological structure. The nodes define the beginning and ending of an arc, therefore providing the arc with a direction. The arc node list stores this relationship between arcs and nodes in a table. Common node numbers indicate an intersection between arcs. This shown in the example of Figure 5: Arc Node topology .



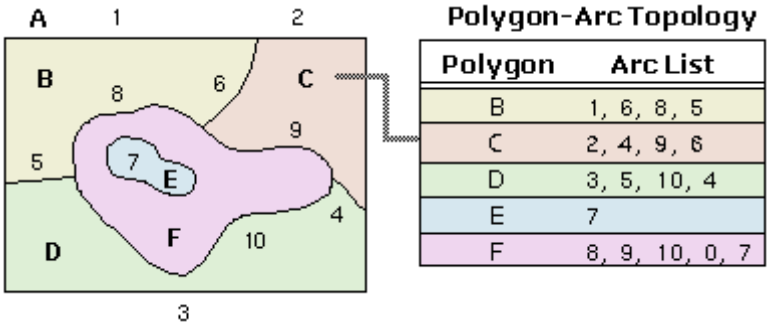
Arc	From Node	To Node
1	10	11
2	11	12
3	11	13
4	13	16
5	13	14
6	14	15
7	14	17

Figure 5: Arc Node topology (ESRI, 2010c)

b) **Planar partition:** the principle of subdividing a plane into different polygons. defining an area by one or more boundaries. It is applied in two variations in the coverage, shown in figure 5 and 6: Area definition and contiguity.

Area Definition

The difference between the arc node structure that defines a polygon area and the polygon definition in a shape file or feature class is the storage: in a coverage, the arc-node structure represents polygons as an ordered list of arcs rather than a closed arc of x,y coordinates. This is called polygon-arc topology. In the illustration below, polygon F is made up of arcs 8, 9, 10, and 7 (the 0 before the 7 indicates that this arc creates an island in the polygon). Each arc appears in two polygons (in the illustration below, arc 6 appears in the list for polygons B and C). Since the polygon is a list of arcs defining its boundary, arc coordinates are stored only once, thereby reducing the amount of data and ensuring that the boundaries of adjacent polygons don't overlap.

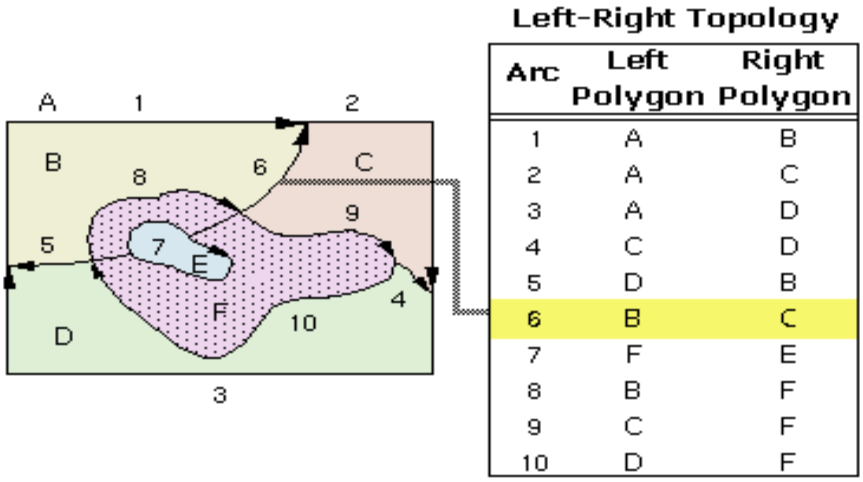


Polygon	Arc List
B	1, 6, 8, 5
C	2, 4, 9, 6
D	3, 5, 10, 4
E	7
F	8, 9, 10, 0, 7

Figure 6: Area definition by arc list (ESRI, 2010c)

Contiguity

Contiguity defines adjacency of polygons by sharing the same arc. The direction of the arc that is also determined helps to differentiate between left and right adjacency.



Left-Right Topology

Arc	Left Polygon	Right Polygon
1	A	B
2	A	C
3	A	D
4	C	D
5	D	B
6	B	C
7	F	E
8	B	F
9	C	F
10	D	F

Figure 7: Contiguity in coverages (ESRI, 2010c)

2.2.2 Feature classes and file geodatabase

With the development of desktop, a data format was needed that could display very quickly and be interoperable with other GIS products : the shapefile. This is a non-topological file format that is still widely used because it is interoperable with other GIS packages. It is very seldom used at Statistics Netherlands, therefore it will not be explained further. After the development of the shapefile, the need for better data management lead to the development of the geodatabase. Within ArcGIS three types of geodatabase are offered:

- The personal geodatabase: first ArcGIS geodatabase format, for single-user use, data are stored in Microsoft Access, maximum storage is 2 GB.
- The file geodatabase: a directory of different datasets like feature classes, feature datasets, raster datasets, attribute tables, maximum default storage is 1 TB, but can be extended to a maximum of 256 TB. The configuration keywords can be used as parameters to improve storage and performance of the data (Childs, 2009).
- The enterprise geodatabase (multi-user geodatabase stored in a spatial database management system such as Oracle, Informix, DB2, PostgreSQL, maximum storage is dependent on DBMS limits)

With the development of the geodatabase, focus has been on fast retrieval of spatial features and their attributes, while there is still need for explicit topology for a large number of tools, such as the generalization tools (e.g. DISSOLVE, SIMPLIFY LINE) that are dependent on identifying topological relations such as neighbouring features. Therefore, Esri has applied methods to build temporary topology for the non-topological data formats (e.g. file geodatabase) such as a Triangular Irregular Network (TIN) (Lee & Hardy, 2006).

2.3 Definition of performance and performance elements

Performance is a non-functional requirement such as security, availability, accessibility, reliability, stability and scalability. These requirements refer to certain qualities that the application or system has to fulfil, whereas functional requirements refer to functionality. The term “performance” within the area of ICT alone has different definitions and interpretations. It is necessary to agree upon one definition which makes development of the benchmark as well as setting up the metrics easier. Moreover, it is up to the GIS users at Statistics Netherlands, how they define performance and the level of which performance level they need.

In technical documentation as well as scientific literature, different interpretations can be observed. ESRI, for example, defines performance as “the speed at which a given operation occurs, e.g. request response time

measured in seconds“ (Pizzi, 2013).This is especially important for the end-user of an application: the higher the speed, the better. But within the time metric, it is very user dependent which time value means that the process is performing well: For example, a response time of 1 minute would be qualified positively for a batch operation, but negatively for a web trans action (Godfrind, 2008). This user dependency also shows the need for clear SLA’s (service level agreements) between user and system manager/IT department. For the user, quality of the data is another important aspect: if the execution time is fast, but the results vary with each identical run or show faults, the performance is still unsatisfactory. For the system or application manager on the other hand, other aspects are related to performance: How efficient are hardware resources used, such as CPU and memory? Does the process lose time because of queuing or because of inefficient memory use? From a system or application managers point of view, the cost is equally important (Godfrind, 2008). Scalability is another non-functional requirement and strongly related to performance: Maintaining an acceptable performance level while increasing the load (workload) on the system.

Surprisingly,Wikipedia (2015) actually provides a definition that is more complete because it integrates, more or less, both point-of –views: it states that “computer performance is characterized by the amount of useful work accomplished by a computer system or computer network compared to the time and resources used.” Here, the time is only one aspect of performance: the percentage of the time used for the actual task counts – a high percentage of the time spent on overhead (for example waiting time between virtual and main memory) means a lower performance. With the expression “useful work” a quality aspect is also introduced: The results of the task have to be of high quality, or high enough to be used. The expressions “useful work” and “workload” and “resources” describe important aspects of performance: The meaning of workload in literature and the real workload at Spatial Statistics has been covered extensively in chapter 2. The workload makes use of various resources, e.g. hardware, software and network capabilities. The state or attributes of these resources (e.g. Clockspeed and number of cores of the CPU) are important parameters for the performance benchmark, whereas metrics on the use of these resources are important to be included in the benchmark.

Resources can be generally described as “system elements that offer services required by other system elements” (Woodside et al., 2007), for example the workload. The resources of Statistics Netherlands, especially of the Spatial Statistics, will be described in chapter 3. In case of Spatial Statistics, for example, different services are required in terms of hardware and software resources than for other teams: higher graphical capabilities, higher calculation power, larger memory.

Performance, in all its different facets and definitions, is the goal of performance engineering, a large domain, but for a long time employed at a later stadium than necessary. (Woodside et al., 2007) provide a fairly complete review of the current state of performance engineering and its elements. The authors analyse a number of problems within that domain. One of these problems is the late detection of performance problems. They propose the following definition:

Software Performance Engineering (SPE) represents the entire collection of software engineering activities and related analyses used throughout the software development cycle, which are directed to meeting performance requirements.

This definition already indicates the complete development cycle, meaning that a large part of performance engineering has already been done by the software developer. How the developer actually designs and carries out the performance engineering is not always clear, which is sometimes the case with commercial developers. Moreover, no role for hardware resources has been depicted in this definition, although performance is also aimed ad use of hardware resources.

2.4 Types of performance benchmarking

Performance benchmarking plays an important role as a decision support on the use of a certain system as well as in research and development (Ray et al., 2011). Paul (2008) describes benchmarking as a "process of evaluating a system against some reference to determine the relative performance of the system." Ray et al. (2011) and Bouckaert et al. (2011) use a similar definition. The reference therefore is a very important element of a benchmark: without reference system, the system under test (SUT) can be measured, but it is difficult to analyse the measurements. With the growth in computing resources (hardware and software) and the use of resources, the need to compare different resources has grown as well, resulting in the development of a large amount of different benchmarks. This also results in a need to classify and standardize performance benchmarks. Benchmarks can be classified based on goal (Klinkenberg, 1997): the qualitative benchmark that should provide information whether system functionality is present, whether it lives up to expectations and whether it is easy to use. A quantitative benchmark tries to answer the question whether the system has the necessary capacity to handle the planned workload. The focus of this research will be on the quantitative benchmark. There are quantitative benchmarks on different levels, such as presented by (Menasce & Almeida, 2001):

- Basic operations that measure only one aspect such as CPU speed with the use of a synthetic workload. Basic operations are also called micro benchmarks.
- Toy benchmarks : also measures on small scale, although with real-life workload.
- Kernel benchmark: These are pieces of code which are extracted from real programs and focus more on the internal resource usage.
- Real program benchmark: In this benchmark, tests full-scale real programs.

The planned benchmark for geoprocessing tools in ArcGIS largely fit into the category of a micro benchmark, although real datasets, a real IT configuration and real software functionality will be used. However, only one tool at a time will be tested. Hennessy and Patterson (2007) assess the real program benchmarks as the most useful, since small scale benchmarks like toy or kernel benchmark can be optimized for one process only and provide a limited view on the system's performance.

For the development of a performance benchmark for spatial databases or GIS systems, few examples are available, let alone standards, although 2 organizations, the System Performance Evaluation Corporation (SPEC) and Transaction Processing Performance Council (TPC) develop and provide performance benchmarks for the computer industry. OGC, the Open Geographic Consortium, is more focussed on conformity, interoperability of spatial data standards, than on performance according to Batcheller et al. (2007). Developing a benchmark for spatial database systems/GI systems is also a challenge because of the complexity of storage and operations that are not suitable for "straightforward" analysis (Batcheller et al., 2007). A number of publications like Ray et al. (2011) and Paton et al. (2000), indicate a lack of standard spatial database benchmarks, although the use of spatial data has grown substantially. Therefore, both publications developed a benchmark for spatial vector data.

Esri has developed a tool to visualize, plan and measure the performance a GIS system, called System Designer and has also published best practices via different media channels and during user and developer conferences. The impression remains that it is mostly directed at larger GIS infrastructures with several tier levels, containing servers, web services and for desktop users (Sakowicz & Pizzi, 2014).

2.5 Performance benchmark development

Because of the lack of standard benchmarks for geographic information systems, a new benchmark has to be developed, tailored for the needs of Spatial Statistics but also suitable for comparable organizations. There are different ways to arrive at a repeatable benchmarking method. Some methods are following a step-by-step, chronological approach, such as shown in Figure 8Figure 8: Performance Testing process : The most important from the beginning is to start with mapping the business objectives and goals. This is followed by

documentation of the non-functional requirements. They are often difficult to define and capture, unlike functional requirements . Derived from the business objectives and processes and a description of the real life workload, the benchmark workload should be defined and documented, followed by definition and documentation of the testing approach and specification of tooling and environment. Based on testing approach and requirements, building the scripts and preparing the data is the following step, followed by validation tests and small scale single user tests. After these tests have been positively executed, the workload can be sized up. If bottlenecks occur, cooperation with the development team is advised to resolve these issues. The concluding step is the analysis of the test results and creation of a performance model.

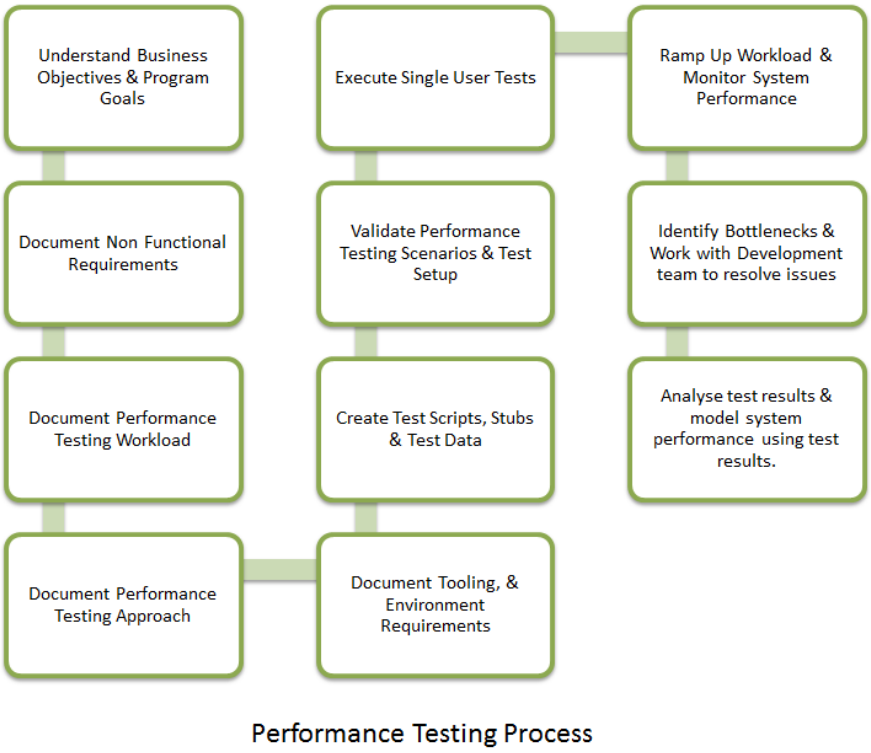


Figure 8: Performance Testing process (Practical Performance Analyst, 2014a)

The approach looks like the waterfall approach in the domain of computer systems design, if indeed all the presented steps have to be followed in chronological order. For the testing environment of SN, a number of steps will not be necessary or return in reduced form or at a different segment of the schema. The documentation of the non-functional requirements will be restricted to documentation of performance only, because this is the foremost issue for the spatial team regarding the possible migration to ArcGIS 10.1 and no other non-functional requirements have not been communicated or documented.

The identification of bottlenecks is only presented as one of the last steps within the process that involves the development team, but should be reoccurring throughout the process: Contact with the IT department, application management and technical staff from the software vendor should be involved at least when defining the testing approach. This could prevent choosing testing parameters that are not viable or meaningful. It should be noted that there is no application manager for ArcGIS at Statistics Netherlands which means that there is no staff member available who can combine knowledge of the geo-ICT infrastructure and new developments of the software with knowledge of user needs.

When evaluating benchmarks the following information needs to be provided in order to interpret the results (Menasce & Almeida, 2001): The benchmark environment (systems configuration, workload, method), the representativeness of the benchmark workload and the properties of the system or new system configuration

the implementation of which has to be decided upon. Therefore, creating the benchmark also introduces the necessity to document the benchmark components very extensively. Menasce & Almeida (2001) assign the following qualities requirements for a successful benchmark:

• Relevance: It must provide meaningful performance measures within a specific problem domain.

• Understandable: The benchmark results should be simple and easy to understand.

• Scalable: The benchmark tests must be applicable to a wide range of systems, in terms of cost, performance, configuration and to be able to cope with a growing size of the input data.

• Acceptable: The benchmarks should present unbiased results that are recognized by users and vendors.

These requirements are not easy to meet, but the acceptability is probably the most difficult: The user group as well as the software developer involved have different stakes within this research project, therefore results could be difficult to accept at least for one of these parties, although the users remain dominant in this process.

Whereas Menasce & Almeida (2001) provide certain qualities, Bouckaert et al. (2011) describe specifications that every benchmark should include:

- A benchmark scenario (description of test settings, tools, data).
- Performance evaluation criteria (describe the high-level focus of the benchmark output).
- Performance evaluation metrics (quantitative measure of a specific quality of a SUT).
- Performance evaluation score (metric value).

The authors also add an important quality to a benchmark: configurability, the test bed should support the needs of the benchmark. In our case, part of the test bed is a real, operational infrastructure (fat clients at Spatial Statistics), part of it more configurable (innovation laboratory), although also dependent on policy, staffing and system requirements.

2.6.Performance workload

The workload can be defined as the demand that is directed at the IT system resources or the database infrastructure: These are e.g. the data and various data parameters such as the amount of datasets, size of the datasets (in records, attributes, memory) and data structures but also actions directed at the resources such as for example queries, transactions or data editing (Mullins, 2010). A thorough analysis of the workload is necessary to be able to develop a performance benchmark that is representative for the organization's information processes. Many literature examples view the workload from a technical, resource related angle. On the other hand, a workload is also seen as a **"**logical set of activities that needs to be performed by users towards achieving a certain (business or customer) goal" (Practical Performance Analyst, 2014b). With that definition, two types of workload can be defined: the business and the infrastructure workload. The business workload encompasses the activities within the IT infrastructure that are directly related to business objectives. For the spatial team, producing Spatial Statistics for a wide range of customers is an important business objective.

The business workload placed on the GIS infrastructure contains, amongst others, the size of the datasets, the type of the geoprocessing tools, the frequency of the use of certain tools, the number of users and the user types. The infrastructure workload relates to the resource utilization patterns of that are the characteristics of a business workload. This could be CPU or I/O behaviour of patterns related to memory use. This is difficult to analyse in our case because of the lack of infrastructure workload data.

It is also possible to differentiate between real and synthetic workload, which is already on the level of benchmark development: Whereas real workload applies real life applications, queries and real data, a

synthetic benchmark uses either a simulation of the resource use of the real workload and/or synthetic data. The advantage of synthetic data is that one can control scaling (Simion et al., 2012) and change in algorithm much better. Moreover, privacy or data ownership issues can be tackled with synthetic data. Irregular data, especially a problem with spatial data, can also blur the true reason of less performance, for example an algorithm that is less efficient. However, the fact that they are not real and mostly very regular in shape and distribution also questions the usability of the results for real life situations. A number of related work examples such as Paton et al. (2000), Simion et al. (2012) use a synthetic workload, whereas Kaler (2012) uses a synthetic workload and a real workload. The synthetic workload is used to detect the direct influence of data structure of spatial data on performance. The goal of a synthetic workload in this case is to "measure performance in a controlled environment", leaving space to the researcher to try different performance parameters.

2.7 Performance resources

Before creating a performance benchmark for the spatial team, resources and workload have to be identified. The resources of the Spatial Team are, of course, dependent on the total IT infrastructure of Statistics Netherlands. The following sections will first deal with the general description of important resources in literature before describing IT infrastructure and resources of the spatial team within the IT infrastructure. A resource can be viewed as a "system element that offers services required by other system elements". Resource and workload interaction results in performance, making the resources a very important set of parameters (Woodside et al., 2007). Resources include:

- Hardware (disk, CPU, bus, I/O and storage, memory, logical resources, network)
- Operating system
- Processing resources (processes, threads)

This section will not provide an exhaustive description and definition of all resource components that are listed but to provide an overview of the most important parts and clarify their relationship to performance. For example, if the I/O time has to be reduced, data that is related (spatially) should be stored within one memory unit as much as possible or have the same entry within an index.

2.7.1 I/O

I/O stands for "input/output", directing at retrieving data from and writing data to the (physical) disk, network or other device (monitor, moveable storage, printer). The disk I/O process is the slowest part of a computer operation and the development of I/O speed is not as fast as development of CPU speed. According to (Hennessy & Patterson, 2007), I/O has been severely neglected in performance optimization. Therefore, reducing I/O to and from disk, can improve performance significantly.

2.7.2 Disk

Different drive or disk technologies have emerged to improve performance and to save power. The "traditional" disk technology, the HDD (Hard Disk Drive) is still used very often. How does the (HDD) disk access its data? The operating system directs the disk through the process which mainly consists of the following stages (University, 2005):

- Rotational Latency/Delay (RT) – the disk head has to rotate to position itself to the right sector on the disk
- Transfer time (TT)- the block of bits has to be transferred, e.g. to memory

These stages are managed by the disk controller. The disk access time is determined by:

Disk Access = seek time + rotational delay + transfer time + controller overhead

In ESRI technical documentation (ESRI, unknown) it is stated that ArcGIS systems are usually CPU bound, but that systems with high volume requests written to a single output are probably more I/O bound. This would imply usage of faster disks, e.g. an SSD or RAM disk to improve performance. An SSD has, unlike the HDD no moving mechanical components. It has an integrated circuit which results in faster access time to data on any location on the disk.

2.7.3 CPU (Central Processing Unit)

The CPU is the central processing unit. Computer calculations are CPU bound if the "rate at which process progresses is limited by the speed of the CPU" (Stackoverflow, 2009). The most important properties of a CPU to take into account are processor type, clock speed (measured in Ghz, relating to the number of cycles per second), cache size and number of cores (BBC, 2015). Simion et al. (2012) have researched the behaviour of spatial calculations regarding its resource utilization. A remarkable result is that 45% of the execution time on average is based on CPU stalls. This happens due to a gap between the CPU speed and the different levels of memory speeds. The growth of data and of analysis and data complexity could lead to an even wider gap.

2.7.4 GPU (Graphics Processing Unit)

The GPU has been original developed originally to accelerate graphics. The acceleration is achieved by using thousands of parallel cores. This mechanism is used more and more for heavy workloads, like geoprocessing on a large scale. There have been several examples of research for the application of GPU which will be further described in

2.7.5 Memory

Memory is constructed as a hierarchy, moving from high level performance but small space down to low level performance but large space for storage. This illustrated in Figure 9:

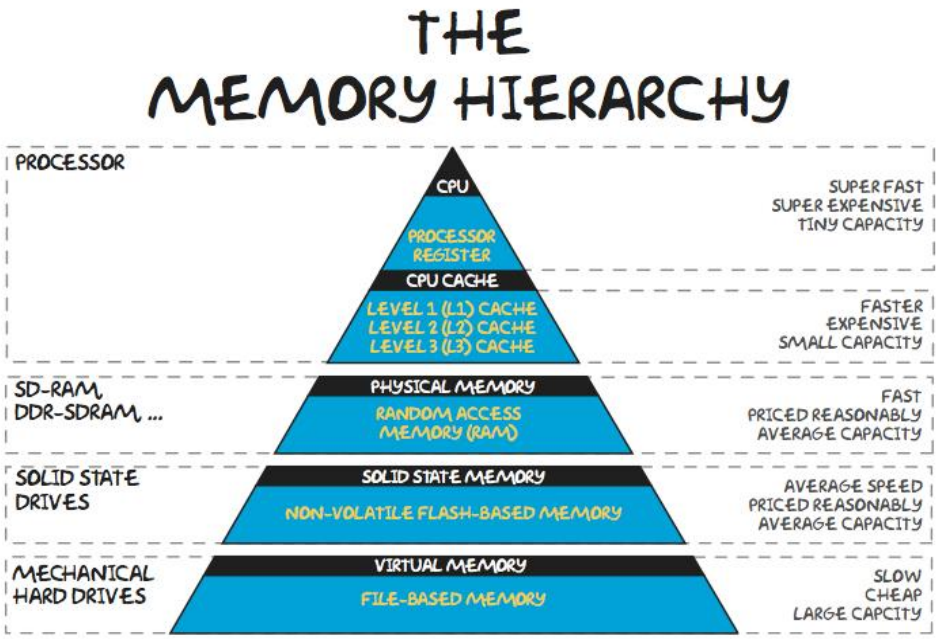


Figure 9: Memory Hierarchy (Teachbook, 2012)

The processes which involves fetching from and storing data in different memory levels can show a number of pitfalls. For example, if requested data are not found in the cache by the processor, a cache miss occurs. The process involving the requested data is stalled temporarily (Hennessy & Patterson, 2007). The same can happen for retrieval of data in other memory components. A pitfall that can result in considerable overhead is a “page fault”. A page fault can occur if a page that is supposed to be accessed from a certain place in memory is actually at another memory location (soft page fault) or even on in the virtual memory on disk (hard page fault). As expected, especially hard page faults can affect performance negatively. A pitfall of another category is a memory leak. These leaks occur if a certain memory space keeps data it no longer needs, which can lead to severe performance deterioration. According to ArcGIS desktop users at Statistics Netherlands, but also on diverse social networking platforms, this problem occurs regularly in desktop and the ArcPy library (Stackexchange, 2011-2014).

Cache

Cache is spaces of memory where data are stored temporarily for quick access during computer calculations. Cache can be found at all memory storage units of the computer: hard drives, hard drive controllers, network cards and especially the processor: the CPU contains 3 different levels of memory types, such as also shown in Figure 9. There is an important factor for the benchmark testing procedures related to cache: Is the tool run directly after starting the application, so that no data are available in the cache (cold) or has the application already been working with the same data, leading to available data in the cache. Ray et al. (2011) include a warm up run. Mostly, a combination of both is observed.

RAM

Random Access Memory, which is stored physically in special DRAM modules on the computer and used to access and store data in a random way instead of sequentially and is therefore suitable to be used for temporary storage. Insufficient RAM can therefore have a negative effect on performance. Geoprocessing performs at its best if carried out within the RAM, the physical memory (Pardy & Hartling, 2013). Therefore, for processing of large datasets, RAM/virtual memory is viewed as instrumental for good performance of overlay geoprocessing. Factors that reduce memory, such as other application processes or, in the case of overlay tools, multiprocessing will negatively affect performance (Hartling, 2012). The limits of RAM are dependent on hardware and Operating System.

Virtual memory

Virtual memory is the space on the hard disk that simulates as main memory for data that should be in the memory but does not fit; therefore it is rather a memory management method. Virtual memory can be larger than the physical memory (Bell, 2013). With this "workaround", limits of physical memory can be dealt with, and reduce I/O activity for memory swapping to and from RAM (Bell, 2013). Still, the I/O from and to the hard disk is very slow, as shown in Figure 9 compared to processor memory.

2.7.6 Operating System

Except for ArcGIS Pro, all ESRI software has been designed for 32-bit, which means that the software is not able to profit from 64-bit resources when installed on 64-bit machines. Since ArcGIS 10.1, Servicepack 1, however, Large Address Aware and 64 bit background processing have enabled the software to profit from 64 bit memory. Large Address Aware (LAA) is already installed with ArcGIS desktop 10.1 and makes an application handle addresses larger than 2 GB. 64-bit is available for background (off screen) processing via an additional installation on top of ArcGIS 10.1, Servicepack 1. Unfortunately, this installation is not available at SN. According to ESRI, 64 background processing does not directly answer performance problems, but is more scalable to process large datasets. Outside ArcGIS, 64 bit processing can be applied with ArcObjects or Python (Pardy & Hartling, 2013) The 32bit version of Windows can access up to 3Gb of physical memory with the user memory setting increased (Hartling, 2012). For Windows 64bit up to 192GB of physical memory can be accessed (Hartling, 2012). A 32bit application, for example ArcGIS Desktop, can access almost twice as much memory when run on the 64bit operating system (due to being Large Address Aware). What needs to be considered when running the geoprocessing tools via a Python script is whether Python has been installed with the 32 bit version?

2.8 Metrics

Performance benchmarking requires "performance measurement". For this purpose, an abundance of metrics is available, related to time, quality, and resource usage. It is important to choose the metric, or a combination of metrics that is understandable (in order to be interpreted correctly), implementable (the tools to measure have to be available) and provides the clearest view on performance. Which metrics to choose depends on the perspective (end users' perspective or system analyst/performance manager). A well-known challenge of performance measurement is the quality of the measurement: Time, for example, is a continuous variable (Wilson, 2010). Consequently, some kind of error will be included during its measurement and analysis. Therefore, a test scenario has to be run several times for statistical analysis of the measurements. Measuring performance of geoprocessing tools is especially challenging, compared to, for example, measuring performance of order transactions.

2.8.1 Time

Time is still the most common interval measure in performance testing (Wilson, 2010). It can be the metric of different processes: the execution time, also called "wall clock time", which is actually the time most users are interested in. Wall clock time is in fact the time span which also includes other time related processes like I/O waiting time or operating system overhead. In fact, as a rule of thumb the following time spans are included in

wall clock time/execution time: CPU time, i/o time, interpreter start up time and bytecode compilation time (Stackoverflow, 2012). Execution or wall clock time is used in many literature examples, e.g. in (Simion et al., 2012) or (Abdelguerfi et al., 2005) where it is even used as the only metric.

2.8.2 CPU

New technology has improved CPU performance but also complicated measuring CPU use because of the possibilities of multiple cores per processor, hyper threading, shared caches and Power Performance Management that enables dynamic adjustment of CPU capacity dependent on the current workload of the system (Performance Team 2009). Different metrics can be used for CPU:The performance is often indicated in FLOPS (floating point operations per second) or MIPS (millions of instructions per second), whereas the usage is measured in percentage or total time (or time per CPU or core). CPU is often measured as CPU usage in %, which basically means the percentage of time which is used for processing for the total CPU facilities (Rodola 2014). The time that a process has been in CPU is also measured in seconds, subdivided into at least different phases: “User” is the amount of CPU time that is spent on actually on executing the process, “system” is the time spent in the kernel (Stackoverflow, 2012), e.g. to manage I/O tasks.

2.8.3 I/O

The metrics used for I/O are the number of read and write processes as well as the memory size of read and write processes. I/O count as a metric is for example, used by (Arge, Hinrichs et al. 2002). Another I/O metric is I/O time.

2.8.4 Memory

As memory is available on many levels (disk, RAM, virtual, CPU), and is being managed by the operating system in order to fulfil the tasks required, a number of different memory related metrics exist. A number of them apply for all memory levels (RAM, virtual, swap):

a) total available memory (e.g. in bytes)
b) amount or percentage of used memory
c) amount or percentage of free memory
d) amount or percentage read or write of memory

Another type of memory related metrics describes the process of the operating system to assign pieces of memory outside RAM to a certain process (paging). These metrics can express the size, but also the number of so-called “page faults”: processes where the address of the data in RAM cannot be retrieved ad have to be searched in the virtual memory.

2.8.5 Number of observations and statistics of observed metrics

One observation of a given metric of a test scenario is often insufficient to draw conclusions. Influence of system resources can vary per run. In most literature examples more than one test runs are used as observation. Tijssen, Quak, and van Oosterom (2012) uses 5 test runs, whereas Ray et al. (2011) implements 3 iterations. To evaluate the metrics of a series of observations, a number of statistical values can be calculated. Mean or median are often used as a statistic measure. The suitability for this metric is dependent on the measured values. If outliers are observed, this is less advisable because they can lead to an incorrect mean value. Still, it is used in a number of examples: the spatial benchmarks Jackpine (Ray et al., 2011) and VESPA (Paton et al., 2000). The benchmark of Tijssen et al. (2012) uses the mean of the remaining out of 5 test runs (after removing the fastest and slowest response time). The median is the middle value in an ordered set of values when there is an odd number of values and the average of the middle two values when there is an even number of values. The median is less sensitive to extreme values than the mean and is usually a better measure of location than the mean for asymmetric distributions. The mode is the most frequent value within a test set, but is not used very often (Wilson, 2010).

Chapter 3: Performance Bottlenecks at Statistics Netherlands

This chapter is dedicated to research question 1:

Which technical as well as organizational performance bottlenecks can be identified during geo-information processes at Statistics Netherlands and how are they currently dealt with?

To answer this question, a number of issues have to be covered: the GIS applications that are used (section 3.1), an analysis of the workload of Spatial Statistics (section 3.2) and an analysis of the IT resources of Statistics Netherlands and Spatial Statistics (section 3.3). Based on this information, a number of technical and organizational performance bottlenecks can be deducted. They are presented in section 3.4.

3.1 GIS applications

The focus of the research is on performance of geoprocessing at team Spatial Statistics in ArcInfo Workstation versus ArcGIS Desktop. However, the final result of this research needs to lead to conclusions and recommendations that fit within working processes of the entire organization. Therefore, a short analysis of the GIS application landscape will be provided in this section.

Examining that application landscape more closely reveals many different applications, each with its own user group (if known) and type of use. The use of IT resources also varies between uses via thin or fat client (see also 3.3.2 IT resources of Statistics Netherlands). A number of GIS related working processes are based on old software. More recent products like ArcGIS Desktop 10.1 and R Spatial but have not yet found their way as main applications within the working processes. ArcGIS Desktop 10.1 is used only for new or custom products and the BBG, because the rest of the standard products of Spatial Statistics are based on ArcInfo Workstation. R Spatial is used at a team that deals with statistical methods and has been working with the statistical language R for a while, before starting to use its spatial extension. At the moment, this team is using R spatial to map traffic intensities based on data captured by sensors, which are lage, spatio-temporal datasets.

A survey has been held to assess the use of ArcInfo Workstation, ArcView and PX-Map. Table 2: Overview of GIS applications at Statistics Netherlands is based on the report of and communication with staff members of Spatial Statistics. The table also shows that 2 products are not from ESRI: PX-map and R Spatial. While PX-Map has been included in the survey conducted by Spatial Statistics, the use of R Spatial has not been measured. Information on that product and its functionality and performance has not been shared with other teams in a structured way. As explained in the thesis plan, the majority of the Spatial Statistics products are produced with ArcInfo Workstation, but are easy in use only for the GIS staff members that are very familiar with it. The version used on the new fat clients is ArcInfo/Workstation 10.0, which is the last planned release, issued together with ArcGIS Desktop 10.0. The ArcInfo Workstation product life cycle (ESRI, 2014) states that ArcInfo Workstation can still be used together with higher ArcGIS Desktop versions, although support will be retired as of 31.12.2015.

Out of the current ArcInfo Workstation users at SN, there is one staff member with expert knowledge of ArcInfo Workstation and AML, and 3-4 users with high knowledge. These users conduct extensive geoprocessing activities with large datasets, based on AML scripts. The AML scripts are used for the products: Financiële Verhoudingswet, Nabijheidstatistieken (Proximity Statistics), Wijk- en Buurtkaart (Municipality-District- and Neighbourhood statistics), Vierkantstatistieken (Grid-based statistics) and Bevolkingskernen (Urban agglomeration). While users are satisfied with performance and functionality of this application, some datasets have difficulties to be processed, especially the key registers, because of the coverage data limitations. The BAG (Key Register of Addresses and Buildings) e.g. has to be subdivided before processing.

GIS system	Programming /Scripting language	Sectors/teams	Thin/fat client	Use	Number of users
ArcInfo Workstation	AML	Environmental, Energy and Spatial Statistics (SLO): Spatial Statistics	Fat	Extensive geoprocessing	4
ArcView 3.x	Avenue	• SLO • SAL • SDI • SES	2 fat, rest thin	Data visualization, simple analysis, for small part extensive geoprocessing	17
ArcGIS Desktop 9.3/9.3.1 (through December 2014, replaced by ArcGIS 10.2.1 in 2015)	Python, VBA	Throughout whole organization	Thin	Unknown	App. 30
ArcGIS Desktop 10.1	Python, ArcObjects	Environmental, Energy and Spatial Statistics (SLO):Spatial Statistics	Fat	• Data editing • Extensive geoprocessing	11
PX-Map	Developed in C#, .Net, JavaScript but no scripting language to extend application (?)	• SLO • BIM • EBD, EBH • SAL • SDI • SES • Quaternary sector statistics	Thin	Data visualization	20
R spatial	R	Process Development and Methodology (PPM)	Unknown	Statistical analysis of spatial data	Unknown

Table 2: Overview of GIS applications at Statistics Netherlands

As shown in table 2, there are 17 users of ArcView at different sectors of SN. More than 50% uses ArcView monthly or even more often, mainly to visualize spatial data, cartography, spatial check of data or to join tabular data to the map. More intensive use like e.g. a spatial join and more extensive editing of maps is only done by 2 users. A majority of the users is satisfied with ArcView, including the customization that has been developed with the scripting language Avenue (Goedhuys, 2014). A number of production processes are carried out with ArcView: creation of shapefiles of district and neighbourhoods, provinces, delivery of x and y coordinates of different boundary files for the Statline table, delivery of maps for PX-map, distances of nature areas and the visualisation of safety per administrative unit (Goedhuys, 2014).

The use of ArcGIS desktop 9.3/9.3.1 at SN is not documented yet. At the start of the research, ArcGIS 9.3 was used with the same purpose as ArcView via the virtual desktop. There are no reported problems concerning functionality and performance, although one has to consider that this application has not been included in the survey. Since 2015 the ArcView 9.3 licenses have been replaced with ArcView 10.2.2 licenses, but at this stage, it is too early to report about user experiences.

The new fat clients are provided with ArcGIS Desktop 10.1. ArcObjects is the development environment of ArcGIS containing the functionality that can be extended on low level in the programming languages Java and C#. However, the current scripting language is Python (the ArcGIS Python library ArcPy), so most of the new custom products made by the GIS staff, has been developed with Python. Since ArcGIS desktop 10, more possibilities to enhance performance are provided, such as support for multiprocessing as well as background 64-bit processing (ESRI, 2013). Background 64-bit processing has to be installed separately on top of the 10.1 installation. These possibilities are known of at Statistics Netherlands but time to experiment with this functionality has been too limited. Three users use ArcGIS 10.1 for data editing and light analysis, 3-4 staff members have a very high proficiency in ArcGIS desktop (geo)processing tooling and Python scripting and use the application for extensive geoprocessing. Python is a runtime interpreted scripting language as well.

PX-Map is a map visualisation application developed for Norway Statistics by Geodata AS. The light, inexperienced users can visualize csv data easily on the map. The user manual of PX-Map explicitly states that it is not a tool for analysing statistical data. It is to be used only as a helping tool to present statistical data as a thematic map (Statistics, 2009).PX-Map is used by more sectors than ArcView by 20 users, although its use is much less frequent: only 10% use it monthly or more often, whereas 90 % use it a few times per year or even less. The user satisfaction is less positive as well, more than 40% is not satisfied.

R can be used for mathematical and statistical analysis of spatial data (such as for example kriging) as well as GIS tasks like buffer analysis and overlay operations. An example of the result of a geostatistical analysis with R spatial is shown in **Fout! Verwijzingsbron niet gevonden.**. The use of R spatial at Statistics Netherlands has not been documented (or documentation is not available at the moment), relevant information on user frequency and type of use still has to be retrieved.

3.2 Workload Analysis

While the previous sections have addressed the total view of GIS at Statistics Netherlands and the role of Spatial Statistics, this section will focus more on Spatial Statistics and the workload that is making use of the technical resources. Workload is a specific term used in performance engineering. Therefore, the following section will first present a small overview of workload analysis in literature. The products that have been previously described in chapter 1 are delivered after a series of geoprocessing steps, executed from an AML script, or in some cases, a Python script. Figure 10 shows an excerpt of a Python script to determine proximity of park entrances via foot and/or cycle paths. It contains a number of selections on attributes via FEATURE CLASS TO FEATURE CLASS, buffer analysis, aggregation of features via DISSOLVE, overlay operations such as INTERSECT and UNION and transformation of geometry type such as POLYGON to LINE.

```
# Process: Feature Class to Feature Class
arcpy.FeatureClassToFeatureClass_conversion(BRT_WEGDEEL_HARTLIJN, nl, "hartlijn_minr", "\"HOOFDVERKEERSGEBRUIK\" <> 'snelverkeer'", "", "")

# Process: Buffer
arcpy.Buffer_analysis(hartlijn_minr, hartlijn_minr_buffer_v, "1 Meters", "FULL", "ROUND", "NONE", "")

# Process: Feature Class to Feature Class (2)
arcpy.FeatureClassToFeatureClass_conversion(BRT_WEGDEEL_VLAK, nl, "wegdeel_minr", "\"HOOFDVERKEERSGEBRUIK\" NOT LIKE 'snelverkeer%' and \"HOOFDVERKEERSGEBRUIK\" NOT LIKE '\

# Process: Repair Geometry
arcpy.RepairGeometry_management(wegdeel_minr, "DELETE_NULL")

# Process: Dissolve (2)
arcpy.Dissolve_management(wegdeel_minr__2_, wegdeel_sel_dis_v, "", "", "MULTI_PART", "DISSOLVE_LINES")

# Process: Union
arcpy.Union_analysis("E:\\marijn\\test_netwerk_top10nl.gdb\\nl\\hartlijn_minr_buffer_v #;E:\\marijn\\test_netwerk_top10nl.gdb\\nl\\wegdeel_sel_dis_v #", union_wegvak_hartl

# Process: Dissolve
arcpy.Dissolve_management(union_wegvak_hartlijn_buf_v, wegen_dis_v, "", "", "MULTI_PART", "DISSOLVE_LINES")

# Process: Feature Class to Feature Class (3)
arcpy.FeatureClassToFeatureClass_conversion(BBG2010, nl, "prk", "\"BG2010A\" = 40 or \"BG2010B\"= 40", "", "")

# Process: Intersect
arcpy.Intersect_analysis("E:\\marijn\\test_netwerk_top10nl.gdb\\nl\\wegen_dis_v #;E:\\marijn\\test_netwerk_top10nl.gdb\\nl\\prk #", intersect_wegen_park_v, "ONLY_FID", "",

# Process: Polygon To Line
arcpy.PolygonToLine_management(intersect_wegen_park_v, intersect_wegen_park_l, "IGNORE_NEIGHBORS")

# Process: Split Line At Vertices
arcpy.SplitLine_management(intersect_wegen_park_l, intersect_wegen_park_split_l)

# Process: Make Feature Layer
arcpy.MakeFeatureLayer_management(intersect_wegen_park_split_l, intersect_wegen_park_split_l__2_, "", "", "OBJECTID OBJECTID VISIBLE NONE;FID_wegen_dis_v FID_wegen_dis_v VI

# Process: Select Layer By Location
arcpy.SelectLayerByLocation_management(intersect_wegen_park_split_l__2_, "CROSSED_BY_THE_OUTLINE_OF", wegen_dis_v, "", "NEW_SELECTION")

# Process: Copy Features
arcpy.CopyFeatures_management(intersect_wegen_park_split_l__4_, result, "", "0", "0", "0")

# Process: Unsplit Line
arcpy.UnsplitLine_management(result, result_unsplit_lines, "", "")
```

Figure 10: Example of combination of processing tools in script (Zuurmond, 2013)

Performance benchmarking for a complete geoprocessing script is not very reliable because the methodology of the script will be subject to change as well, leaving the single geoprocessing tool as the smallest and clearest unit to perform the benchmark, similar to the “micro benchmark workload” applied by Ray et al. (2011). As shown in Figure 10, a product of the spatial team is a result of a chain of processing tools, which does not only contain geoprocessing tools, but also a number of administrative processing on the tabular data which serve as preparatory steps and stand for at least 50% of the processing tools. At the beginning of the production process, the input data have to be prepared for use, e.g. by discarding attributes that are not used at SN, or geocoding addresses by joining the residential objects with the addresses. A brainstorm session has been held with the “heavy GIS users” to discuss the tools that have to be included. The selected tools have

been prioritized for performance evaluation for various reasons: the tools are often used, performance is problematic in ArcGIS desktop, performance is problematic in ArcInfo Workstation, stability is less in either

desktop or Workstation. The following tools have been prioritized by the heavy GIS users. The following is the FIRST prioritization step:

1. Geoprocessing tools
    a. UNION: instability in ArcInfo Workstation with large datasets, but also slower performance in ArcGIS desktop, frequently used
    b. DISSOLVE: Frequently used, slower performance in ArcGIS Desktop
2. Administrative processing tool
    a. ADMINISTRATIVE JOIN: used many times, big difference in performance noticed between ArcInfo Workstation and ArcGIS Desktop 10.1 (20 minutes in Workstation, 3 hours in Desktop)
    b. UPDATE CURSOR: long execution time, big difference between performance on Windows XP fat client and Windows 7 fat client (Windows 7 performance is slower)
    c. SUMMARY STATISTICS: frequently used tool
    d. FREQUENCY: frequently used tool

As a following step, tests with Workstation have been conducted, also, to test the difference before and after migration of the fat clients to Windows 7 (including migration from Workstation 9 to 10). The test results will be presented in chapter 5. To keep the focus of the research on geoprocessing, the INTERSECT will be added to the workload. Additionally, a tool that has not been mentioned during the discussion but plays an important role within the transition from Workstation to desktop is the NEAR and the OD Matrix, used for the proximity statistics. The NEAR will therefore be included in the benchmark workload. As a result, the DEFINITE set of tools for the benchmark will be as follows:

1. Geoprocessing tools
    a. UNION
    b. DISSOLVE
    c. INTERSECT
    d. NEAR
2. Administrative processing
    a. ADMINISTRATIVE JOIN (JOIN FIELD)
    b. SUMMARY STATISTICS
    c. FREQUENCY

Due to the scope on geoprocessing, the administrative tools will not be tested with each factor (only in default settings and with a different hardware configuration); the geoprocessing tools will remain the focus of the benchmark to be developed.

3.2.1 Tools

3.2.1.1 INTERSECT

This tool calculates the geometric intersection of the input features. These features can be of different geometry type. The output dataset will contain only features or portions of features which overlap in all layers and/or feature classes, such as shown in Figure 11: The overlay of the two datasets (one containing 3, the other containing 2 objects) results in an output dataset containing 2 objects. Less time is expected to be spent on I/O activity for the output geometry, contrary to the UNION, which has to calculate and write output geometry that includes the new geometry derived from the intersection of both input geometries as well as the geometry that does not intersect. Before starting the actual intersection process, the tool first determines the spatial reference. Then it has to crack the features (inserting vertices at the intersection of feature edges) and cluster (snapping together vertices that are within x and y tolerance). During the next step it analyses the geometric relationships between all feature classes and writes the new features to the output (ESRI, 2013e).

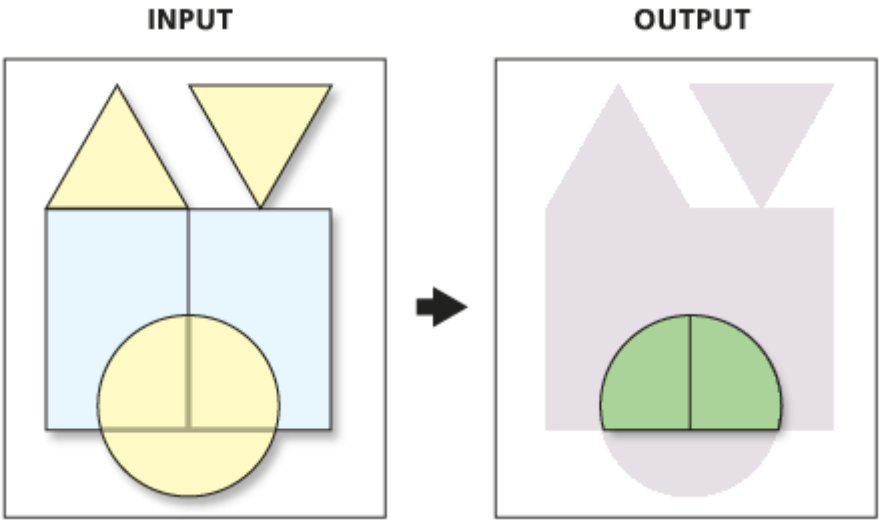


Figure 11: INTERSECT input and output features (ESRI, 2013e)

The usage of resources will be probably similar to the UNION, as it is a comparable tool, yet the output feature class will be smaller in records and memory, therefore one could assume that less I/O capacity would be needed for writing the result to the output.

3.2.1.2 UNION

A frequently used geoprocessing tool, using 2 input feature classes or feature layers and computing the geometric union of all input features and their attributes. Running the UNION with one input dataset is also possible if overlap between the features within this dataset has to be detected. The resulting output feature class will show polygons that represent the geometric union. Figure 12 shows the 3 objects of the input, resulting in 7 output objects derived from the overlay process. Before starting the actual union, the tool first determines the spatial reference. Then it has to crack the features (entering of vertices at the intersection of feature edges) and cluster (snapping together vertices that are within x and y tolerance). During the next step it analyses the topological relationships between all feature classes and writes the new features to the output (ESRI, 2013i). Spatial indexes can be added to input datasets to improve performance. At Spatial Statistics, the default spatial indexes are used.

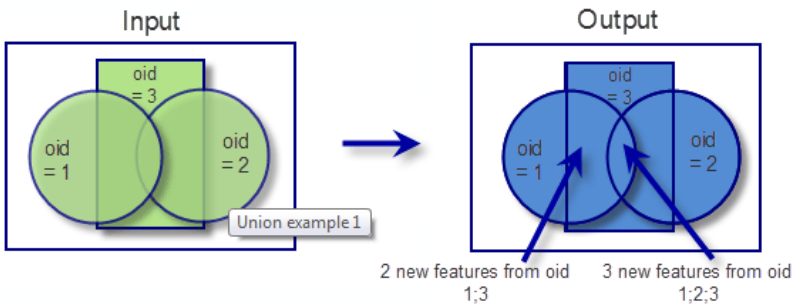


Figure 12: UNION input and output features (ESRI, 2013i)

A UNION is often calculated with the Land Use File (BBG) for two different years, to easily identify changes between the two datasets. This type of historical analysis is also applied to other polygon datasets. Similar to

the INTERSECT, the tool did not work satisfactory with the BBG as a coverage file during the baseline tests. An error message with the phrase “Too many files-FILEGET” was rendered. UNION is a tool that will probably use a lot of I/O (to retrieve the files as well as the indexes) but also memory to load the files as well as CPU to calculate the geometry of the input files.

3.2.1.3 DISSOLVE

This tool is used to aggregate adjacent geometric features, often based on a certain attribute or more attributes. As a next step, the aggregated attributes are often summarized with different statistics. Features with the same value combinations for the specified fields will be aggregated (dissolved) into a single feature or a multi-part feature. It is also possible to dissolve without attribute (dissolve field), e.g. to dissolve overlapping features (ESRI, 2013c). The steps that DISSOLVE has to take are roughly: selection of records with the same attribute value (if this option is used), identification of neighbouring polygons within the resulting subset of records, removing vertices of the shared boundaries, topology check of the resulting polygons and storage of new polygon set. The identification of neighbouring polygons is a step that can be expected to be easier with explicit topology, applied in the coverage data format of ArcInfo Workstation.

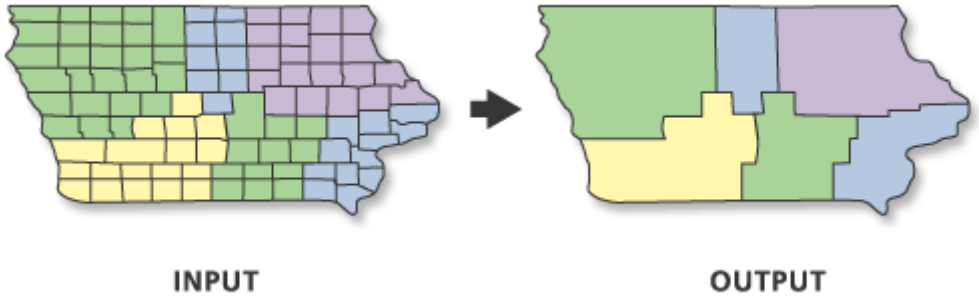


Figure 13: DISSOLVE input and output features (ESRI, 2013c)

3.2.1.4 NEAR

The near is a proximity tool. With this tool, the shortest distance between each feature (point, line, and polygon) and the near feature (point, line, and polygon) is calculated within the search radius according to a set of rules. It determines the distance from each feature in the input features to the nearest feature in the second dataset, within the search radius (ESRI, 2013f). The table of the input dataset will be updated with the feature ID of the nearest feature and the distance from the input feature to the nearest feature.

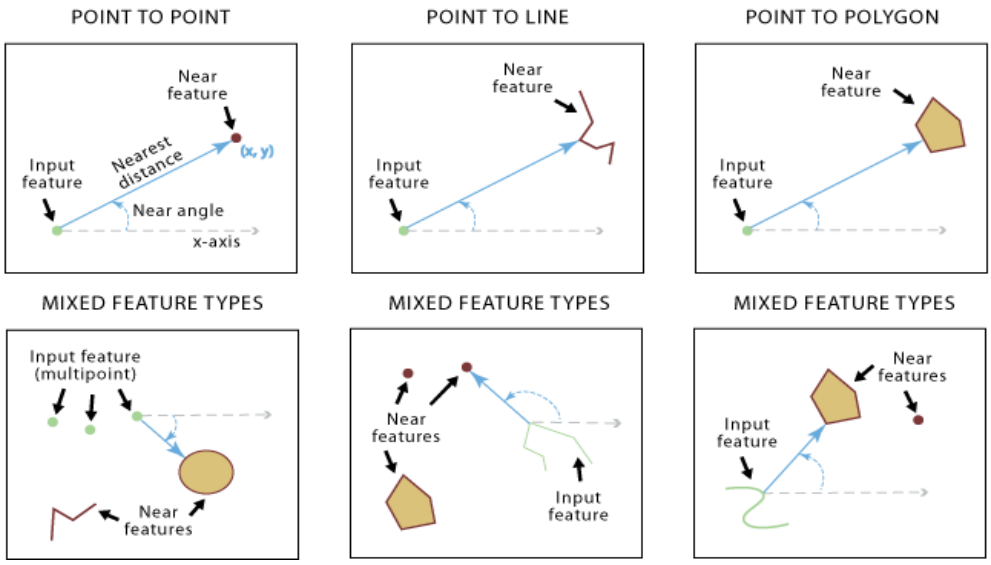


Figure 14: Different NEAR options (ESRI, 2013f)

This tool is stable only in ArcInfo Workstation at SN, using a dataset on national scale. The proximity calculation will put a high load on the CPU use, whereas the part of updating the table will be I/O intensive. Loading the input and near data will also be memory intensive.

3.2.1.5 ADMINISTRATIVE JOIN

The administrative JOIN is applied with the AML command "JOIN ITEM" or the Python command "JOIN FIELD". As explained before, this tool is a very important, frequently used preparatory step that adds attributes of a related table to the source table, enabling further administrative or spatial analysis. Even though it is not a geoprocessing tool, the JOIN is an operation that costs a lot of time for large datasets. The source datasets keeps all attribute fields, the fields from the join table can be selected . JOIN FIELD is preferably used for a 1:1, 1:n or n:1 relationship. JOIN FIELD is often preceded by a SUMMARY STATISTICS (in case of a 1: n relationship, a selection can be made, based on the field statistics). To improve performance, an index is often added based on the join field of the join table. Most probably, more I/O capacity is needed in comparison with CPU capacity, because the index (if present) and all table rows of the join table have to be read and coupled to the corresponding field of the source dataset. Another possibility is the ADD JOIN, which at first creates a virtual join of the tables. A step to create a real table is required, for example with COPY ROWS/COPY FEATURES or TABLE to TABLE/FEATURE CLASS to FEATURE CLASS. The experiences at Statistics Netherlands with this method have been mixed: faster performance has been noticed but less stability.

3.2.1.6 SUMMARY STATISTICS and FREQUENCY

These tools are part of the "Statistics" toolset. The SUMMARY STATISTICS is available in all licenses whereas the FREQUENCY only in the advanced license. The SUMMARY STATISTICS calculates a statistical value (e.g. sum, mean, maximum, count) over a specified field whereas a FREQUENCY counts the occurrences of a value of a specified files and stores the result in a separate output table (ESRI, 2013h) and (ESRI, 2013d).

3.2.2 Data

The geoprocessing activities are carried out on vector datasets, largely on national scale. Some datasets are processed per map sheet or by a smaller administrative unit, e.g. province. Raster data like imagery are used as a reference layer mostly. The BRT and imagery is delivered in map sheets. To conduct a number of tools such as e.g. the NEAR, the data have to be processed in smaller units, for example on municipality level, instead of national level. An overview of the most used datasets will be provided in the appendix. In the following sections, these datasets will be described in short. The metadata of the datasets will be provided in Appendix 2: Description of selection of spatial datasets.)

3.2.2.1 Basis Registratie Adressen en Gebouwen (BAG)

The BAG (the key register of addresses and buildings) contains the following objects: gebouwen (buildings - contains polygon geometry), verblijfsobjecten (residential objects, contains point geometry, has to be located within one or more buildings), ligplaatsen (anchorage, contains polygon geometry), staplaatsen (pitch, contains polygon geometry), nummeraanduiding (house numbers, no geometry), openbare ruimte (street, no geometry), woonplaats (residence, polygon geometry). The polygon geometry of pitches and anchorage is turned into points. The geometric features of the BAG is delivered by the Kadaster in the open format GML (Geography Markup Language) and converted to file geodatabase at SN to be ready for use in ArcGIS. The address objects are administrative but very often joined to the spatial tables. It is a very heavy dataset, varying per year from 8 to 10 GB. The tables contain many attributes, approximately 15 per object. Often, the tables are stripped of unnecessary attributes to improve performance. Processing steps that often occur with these datasets are:

- FEATURE TO FEATURE/TABLE TO TABLE (e.g. to remove or rename attributes)
- FEATURE TO POINT to create centroids of anchorages and pitches (before adding them to the residential objects)

- JOIN FIELD - Administrative join of house numbers to a combined table of residential objects, pitches and anchorages to obtain addresses with coordinates
- SPATIAL JOIN of residential objects per building
- DISSOLVE of building polygons

A practical example of a processing flow is the identification of addresses within a block of buildings (that use block heating) and creating an overview of those building blocks per district or neighbourhood.

- Feature to point to create centroids of pitches and anchorages
- Feature to feature of residential objects to remove/rename field and to keep the original dataset intact
- Merge or append centroids of pitches and anchorages with residential objects (creating a combined dataset, called the VSL (Verblijfsobjecten, Sta- en Ligplaatsen)
- Join field between VSL and house numbers
- Spatial join with district- and neighbourhood boundaries

3.2.2.2 Nationaal Wegenbestand (NWB)

The NWB is a digital, geographic dataset on a scale of 1:10000 that contains almost all roads in the Netherlands. It contains the roads that are maintained by the National public authorities, provinces, municipalities and water boards. The NWB forms the network used for the proximity analyses of SN. The NWB is stored as a coverage, which implies that the proximity analyses are conducted with an AML script. It also means that topological information regarding the network is stored explicitly in the dataset. Statistics Netherlands always prepares the data set with a correction of the drive direction and the connection of orphans with the network. The proximity statistics are calculated via a NEAR from address coordinates to projection points on the network and an Origin Destination (OD) matrix via the network from the projection points to facilities.

3.2.2.3 Bestand Bodemgebruik (BBG)

The BBG is produced by SN, but also re-used by SN and a large number of partners. It is a dataset derived from imagery where areas are assigned with codes depicting a certain type of land use. Processing steps that often occur with this datasets are:

- A union with BBG from two different years to detect changes or mistakes

3.2.2.4 Wijk- en buurtkaart (district and neighbourhood boundaries)

This dataset contains the boundaries of districts and neighbourhoods within a municipality and is used to produce statistics for small areas. Processing steps that often occur with this datasets are:

- A DISSOLVE on neighbourhood-, district-, municipality borders
- Combination with statistics per administrative level

3.3 Performance resources

3.3.1 Location of GIS elements within the Statistics Netherlands Infrastructure

Within the general IT infrastructure of Statistics Netherlands, the spatial team conducts its geoprocessing activities on the local drive of their fat clients, but still has to connect with the data centre in Almere to access the ArcGIS license server and the user profile. The datasets are copied form the network to the local drive. The user profile is stored

and back-upped within the CIFS[4] storage pool. The executable of ArcGIS is installed on the fat clients via SCCM (Schets & Desabandu, 2014). SCCM stands for System Centre Configuration Manager, a tool that is used for tasks such as for example patch management and software distribution (University, unknown). This is visualized in Figure 15. It has to be noted that this figure does only show the elements of the ArcGIS users. The resources and workload of R Spatial has not been investigated in this research project.

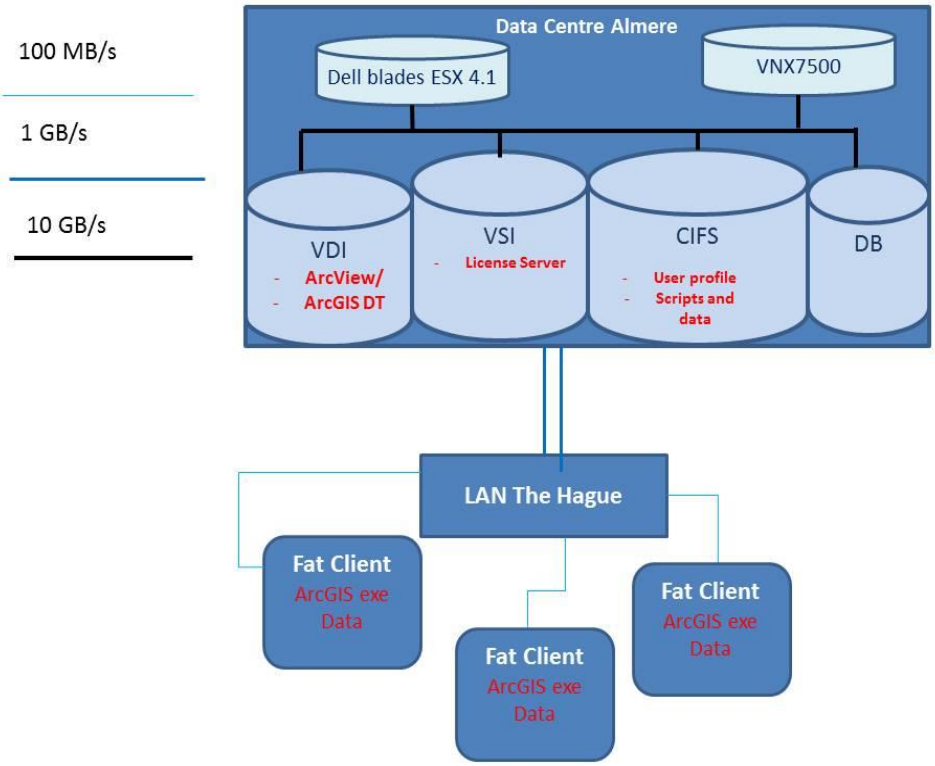


Figure 15: ArcGIS elements within IT infrastructure SN

Figure 15 shows that the fat clients have to use a relatively slow connection of 100 MB/s within the office LAN and a "medium" connection of 1 GB/s for the WAN to connect with the virtual servers of the VDI storage pool to access the license server, or the CIFS servers to access the scripts and copy data from the network drives. The license server has to check if the user possesses the necessary licenses to execute a certain tool (e.g. extension Spatial Analyst or 3D Analyst). The script is accessed and interpreted during the process, meaning that an intermediary language is needed to translate the code into machine understandable language.

It is not sure how many times the activity of reading code lines via the LAN and WAN connection occurs during a geoprocessing tool and thus influences the performance. This could be included as a factor in the benchmark by comparing performance of the geoprocessing tool with the script stored locally and on the network. However, it should also be noted that the network bandwidth will be upgraded during the course of 2015.

The user profile is additionally accessed via the LAN and WAN connection, as it is stored within CIFS virtual server. The support page of ESRI states that "ArcGIS for Desktop relies heavily on the Windows User Profile", which means that a new profile has to be created, once it gets corrupt. If the "My documents" folder is redirected to the data centre, this could cause performance problems for ArcGIS Desktop 10.x, because basic

[4] *Common Internet File System (CIFS) is a protocol that programs use to make requests for files and services on remote computers on the Internet (Rouse, Margaret, 2005).*

files such as the default file geodatabase are stored in that folder. Local storage would be advisable, according to Esri (ESRI, 2012).

3.3.2 IT resources of Statistics Netherlands

Although the GIS staff of Statistics Netherlands conducts most of its geoprocessing activities on the local drive, some elements are linked to the network. Therefore, the description of the total IT infrastructure of SN is useful to understand how the geoprocessing activities interact with the rest of the infrastructure and to identify bottlenecks that affect the working processes at the spatial team. Moreover, knowing the background of the current situation at the spatial team, such as the introduction of virtual desktops as well as the re-introduction of the fat clients helps to evaluate possible solutions towards a sustainable migration of the GIS. Enterprise IT infrastructures of large organizations are often complex and difficult to maintain. This is certainly the case for Statistics Netherlands: it needs an IT infrastructure that can handle a number of different challenges: Storage of very large databases, a very high use of complicated queries on different enterprise databases and a high level of security to protect sensitive micro data. The amount of data storage is extremely high with 1 petabyte (1000 terabyte) as well as the email traffic with more than 7 million emails per month. A very noticeable figure is the percentage of spam email (80-90%). There's a noticeable difference between the number of (virtual) desktops and the number of users (Stormen, 2013). There are more Windows 7 thin clients than staff members for the following reasons: Many desktops are used for acceptation purposes, users with more than one role, and non-active users. Non-active users are cleared several times per year(Schets & Desabandu, 2014). Statistics Netherlands currently has app. 2.200 users of its ICT infrastructure.

Datacentres and network

Geographically, the users are spread over 2 office locations in The Hague and Heerlen, such as shown in Figure 16. The data storage, system transactions and systems security are supported by the backbone of the ICT infrastructure, the data centre in Almere, which is also maintained by Statistics Netherlands itself. A back-up data centre is available at Oude Meer, which is connected only to the data centre in Almere.

Figure 16: Overview Statistics Netherlands location of offices and data centres (Stormen, 2013)

The network within the office location is organized as a LAN (see Figure 17). The connection within the LAN has a capacity of 100 MB/s, which is considered the weakest link within the network. The Data centre in Almere is connected with the LAN's in Heerlen and The Hague via a WAN (Wide Area Network), whereas the two LAN's are connected as well. The Datacentre in Almere has only one access to the internet via a DMZ (de-militarized zone) unit to protect the IT infrastructure from network attacks. The WAN network between the office locations and the data centre has 2 Ethernet connections of 1 Gb/s (Figure 16 and Figure 18) whereas the servers within the data centre that support the (virtual) desktop are connected to storage pools via an Ethernet connection of 10 GB/s.

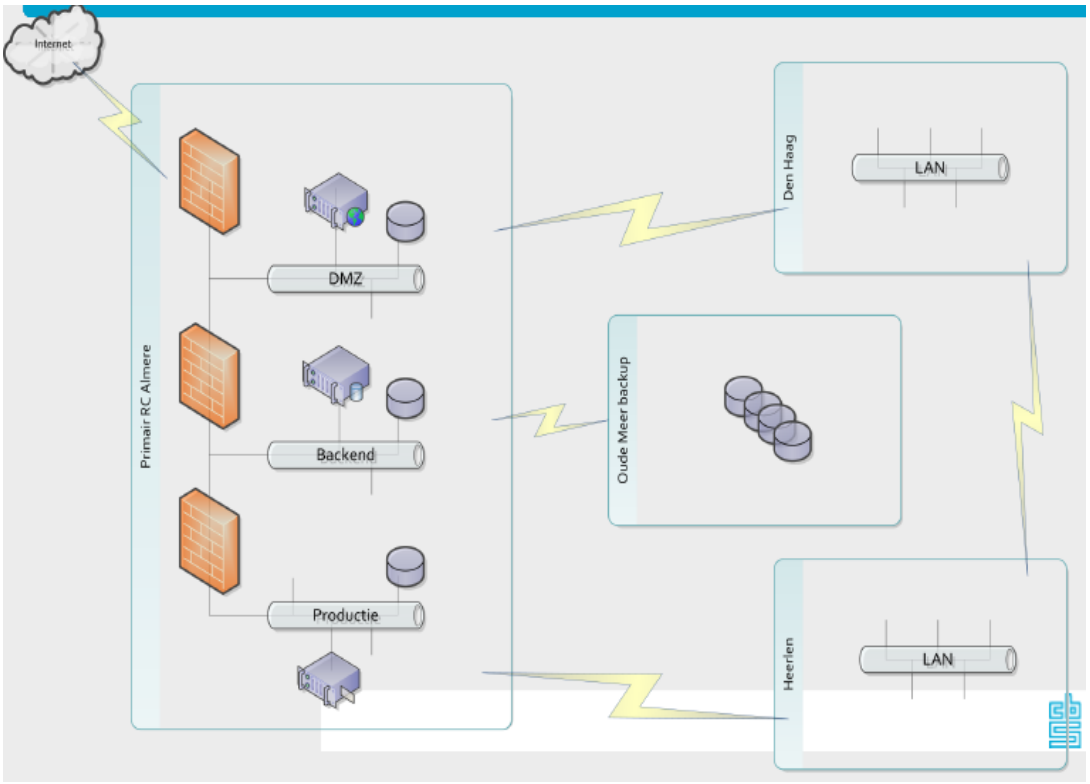


Figure 17: Overview Statistics Netherlands network entities (Stormen, 2013)

Virtual infrastructure

The majority of the users are working via virtual thin clients, using a VDI (virtual desktop infrastructure). VDI enables access to a virtualized desktop, which is hosted on a remote service over the Internet (Techopedia, 2014). The choice for thin clients has been made because of the implementation of a separate data centre at a location in Almere, which is located too far from the offices of Statistics Netherlands to provide a workable transaction time via the WAN between the fat clients and the data centre. There are 4000 Windows 7 thin clients provided via app. 1000 VMWare (currently version 5.5, migration to this version is assumed to have taken place in June 2014) virtual servers (Stormen, 2013), running on Dell blade servers at the data centre in Almere. Those servers are coupled to so-called "storage pools" (connected with Ethernet capacity of 10 GB/sec) which differ in memory capacity (Stormen, 2013) and (Schets & Desabandu, 2014). Figure 18 shows the hardware resources that are used to support the main IT infrastructure:

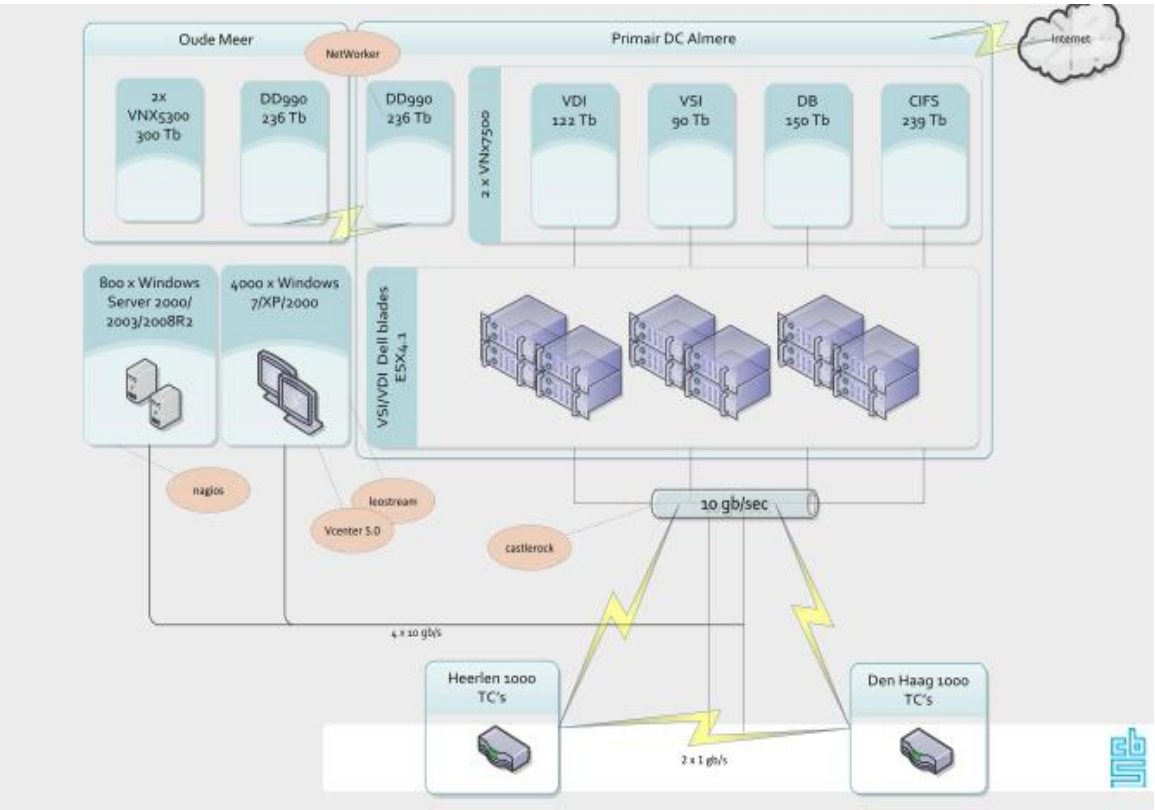


Figure 18: hardware resources (Stormen, 2013)

The storage pools are visualized on the right of the schema, showing the storage pool per function:

Storage Pool	Memory	Function
VDI	122 Tb	Supports the virtual desktops
VSI	90 Tb	Supports the virtual servers in the data centre
CIFS (Common internet file system)	239 Tb	Enables file and service requests. The storage pool supports storage of e.g. user profiles. As well as the use of network drives G:, H: for data storage
iSCSI (internet small computer system interface)	150 Tb	isCSI is a protocol for databases and supports database storage

Table 3: Storage Pools and their function

Of those storage pools, only the virtual desktops are not back-upped, except for their user profiles, because they are stored in the CIFS storage (Schets & Desabandu, 2014).

Virtual desktops vs use of fat clients

The ICT policy of Statistics Netherlands primarily strives to have a thin client workspace for every staff member. However, ArcGIS desktop and ArcInfo Workstation appeared to perform poorly on GIS calculations within the implemented virtual infrastructure, therefore leading to re-introduction of fat clients for the heavy GIS users (Opperdoes, 2014). The use of GIS software via a virtualized desktop has proven to be insufficient for the application of GIS that goes beyond creating and viewing of simple maps, at least for the virtual construction, chosen by Statistics Netherlands. A number of light users (it is not sure how many exactly) work with ArcView 3.1 or ArcGIS Desktop 9.3 via the virtual environment for visualization.

In general, visualization of spatial data, analysis and geoprocessing within a virtualized environment is a viable option. As for ESRI products, ESRI has published test results of a configuration of ArcGIS Server 10.1 Enterprise configuration running on VM VSphere 5.1 virtualization (ESRI/VMWare, 2013). Some of the most important findings where that a virtual configuration with less machines with more CPU cores provide better performance than a virtual configuration with more machines and less CPU cores. The most advantageous virtual configuration revealed a minimum difference with the physical configuration. The described situation is not truly comparable with the situation at Statistics Netherlands; SN has ArcGIS 10.1 desktop as well as ArcInfo Workstation 10 and VMware 4.1. Additionally, the IT infrastructure and workload in the ESRI/VMWare publication are not comparable with SN. Virtualization of the GIS infrastructure has also been implemented at a comparable organization in the Netherlands (as will be described in detail in chapter 7), the Netherlands Environmental Assessment Agency (PBL), physical GIs servers are accessed via virtual desktops, showing satisfactory results (Desabandu, Put, & Spoon, 2014). To explore viable virtualization architectures that would work with the needs of the spatial team would require a much larger research scope. It is, however, worthwhile for the spatial team to consider the current virtual environment at SN, and to find out, what is necessary to make the virtualization technique work for geoprocessing. Several possibilities could emerge:

- A different virtualization software package or a different version.
- Change in resources: ArcGIS servers (physical or virtual) with more capacity.
- A change in the server infrastructure: e.g. use of a physical license and ArcGIS desktop server, instead of a virtual one. This would be comparable to the situation at the PBL.

3.3.2.4 Fat clients properties

Every GIS specialist has been provided with a fat client. At first, 32 bit fat clients with Windows XP operating system with the following properties have been used:

Operating system	Windows XP
Hardware provider	Fujitsu
Windows performance index	Not available
System type	32 bit
RAM	3 GB
Processor type	Intel® Core™ 2 Duo CPU E7300@2.66 GHz 29B
C drive	Physical C drive
Version ArcGIS desktop	9.3.1n (1 test machine also with 10.0)
Version ArcInfo Workstation	9.0 (1 test machine also with WS 10.0)

Table 4: Properties Windows XP fat client

After the migration to Windows 7 (project PUMPS), most of the fat clients were traded for machines with the following properties:

Operating system	Windows 7 Enterprise SP 1
Hardware provider	Fujitsu
Windows performance index	5.2
System type	64 bit
RAM	8 GB (7,85 GB available)
External memory	2 TB
Processor type	Intel® Core™ i5-3570 CPU @3.40 GHz 3.40 GHz (4 cores)
C drive	Part of the C drive (roaming profile and My Documents) reconnects to the datacentre
Version ArcGIS desktop	10.1
Version ArcInfo Workstation	10.0

Table 5: Properties Windows 7 fat client

These fat clients have been provided in concurrence with system requirements for ArcGIS, but do not belong to the category of high performance hardware. Additionally, the hard disk is a mechanical hard disk (HDD), although drives like SSD's (solid state disks) are available on the market, although at much higher cost, and are known to provide much better performance. However, the fat clients provided with Windows 7, already have hardware properties that should lead to higher performance, e.g. the processor that changes from 32 bit to 64 bit. 64-bit processors are expected to process faster than 32 bit because they have more cores: dual core, quad core, and six core for home computing (Computerhope, 2015) and because they support a higher amount of memory (RAM). Not all software is yet programmed for 64-bit. ArcGIS desktop is not designed for 32-bit, it can be expected that ArcInfo Workstation and ArcGIS desktop will not profit from the 64-bit processor.

It is important to note that the account settings are configured to the network for the roaming profile and the My Documents, whereas the Temp folders are placed on the local drive. The disadvantage of these settings is that the ESRI settings can become corrupt, whereas the My Documents folder does not provide enough space for the immediate storage of data in the default file geodatabase (default.gdb). The default file geodatabase is the home location of the spatial content of a map document. It is synchronized with the Current Workspace of Geoprocessing Environments (ESRI, 2012). The location of the default file geodatabase can be adapted, which is also recommended after a "health check" by ESRI NL (Iparraguirre, 2014).

3.3.2.5 Support of the IT department

Discussions and interviews with GIS users (heavy users) of the team show that the cooperation with the IT department is not always easy, partly because of lack in time and manpower capacity on both sides, but also because of the specific challenges of GIS processes that require special hardware resources and specialist knowledge of optimization of (geo) processing of spatial data. One staff member of Spatial Statistics is tasked

with coordinating ICT requests towards the IT department, whereas the IT department provides one or two staff members that have some extra knowledge of GIS software and are tasked with out the installation and minor errors.

The hardware resources have partially been provided with the fat clients, but the specialist knowledge is not present within the IT department. Additionally, the focus of the IT department is on IT security and standardized, efficient maintenance of the infrastructure with less custom infrastructure as possible. This leads to a higher dependency on the software delivering party, in this case ESRI. Specific SLA's for non-functional requirements such as performance, stability, quality, availability, etc. have not been agreed upon with the IT department. The services the IT department provides is limited to license management and solving small hiccups, although it is difficult for the IT department to provide figures on the exact number of users of the diverse GIS applications. The licenses (24 in total) that are in use at the moment are (Desabandu, 2014):

- 8 advanced concurrent - ArcGIS Desktop 10.1/ArcInfo Workstation 10 mainly for fat client users
- 4 basic concurrent
- 12 basic single use

An upgrade to ArcGIS Desktop 10.2 is planned in the near future, having started with the virtual desktop users in April 2015. Spatial Statistics has decided to have ESRI conduct a health check on the current implementation of the current ESRI products within the IT infrastructure. The health check showed that some components can be configured differently, e.g. the licences and location of the user profile and the default geodatabase. The hardware itself has been evaluated as sufficient. The health check report also revealed a number of results for a number of tools such as the UPDATE CURSORS, JOIN ITEM/JOIN FIELD and FEATURE CLASS TO COVERAGE. JOIN ITEM with ArcInfo Workstation/AML showed significant better results than JOIN FIELD with ArcGIS Desktop/Python: 5 (AML) seconds execution time versus 331 seconds (Python). This has been communicated to the ESRI back office. The UPDATE CURSOR shows better results in Desktop/Python, the combination Windows XP/ArcInfo Workstation performs better than Windows 7/ArcInfo Workstation. Finally, the conversion of large datasets to coverage often results in fatal errors, such as "Too many files" or "invalid topology" (Iparraguirre, 2014). The recommendations of the health check mainly address changes in the configuration: installation of the different software components:

- Install home folder and user settings for ESRI products locally on the fat client, not the My Documents folder, which is situated on the network
- Setting the port number and host name for the license manager
- Minimize the impact of the use of bubble (see **Fout! Verwijzingsbron niet gevonden.**) networks for software installation: Install PyScripter and ArcView locally or expand the tasks of the bubble

These recommendations have not yet been implemented and are not part of the benchmark.

New developments in IT resources

During autumn 2014, a project to migrate to MS Windows 7 operating system has finished, leading to apparent loss in systems performance, which has been perceived by users at different departments. The reasons of these observations are currently researched by the IT department, but no clear conclusions have been found at the moment. Probable reasons could be growth in users, new applications running under Windows 7 (Schets & Desabandu, 2014). An upgrade of the operating system could be combined with an application upgrade as well, such as e.g. at the spatial team. Apart from the IT infrastructure that supports the primary processes at the organization, an "innovation laboratory" has been created to facilitate proofs of concept of innovative ideas, without dependency on the regular IT infrastructure. The tools that are available at the innovation lab are partially hardware facilities, such as with laptops and big data computers with the following properties:

Operating system	Windows 7 professional
Hardware provider/type	Fujitsu Celsius M720
Windows performance index	5.2
System type	64 bit
Disk	4x 256 GB Solid State disks
RAM	64 GB f
Processor type	Xeon E5-2640 2.5 GHz 15MB Turbo Boost (16 cores)
Version ArcGIS desktop	10.2, ArcGIS Pro

Table 6: Specification big data computer

One of the big data computers has internet access, the other only to the closed network. Several laptops are available as well for testing objectives. The infrastructure is completely separated from the SN network and maintenance is organized by the staff of the laboratory themselves. The facility is clearly meant for proofs of concept and not for production. Using the facilities directly for production would mean that the team that uses the facilities would have to pay for it, resulting in very high production costs for the statistical product. Due to the fast developments in hardware but also grow of data volume, upgrade of the innovation lab facilities is being considered. Possible solutions that are thought of by SN are decentralized hardware per team/department, aimed at use for hardware, extension of the innovation lab resources or cooperating with the University of Amsterdam on the use of a super computer in datacentre SARA. Other cloud solutions could be the Amazon Cloud[5] or a big data facility for statistical agencies in Ireland (Desabandu & Emons, 2014). Recently, Spatial Statistics has started to conduct experiments with network analysis in the innovation lab on the big data computer, using ArcGIS Pro.

3.4 Summary: performance bottlenecks

Based on the information of sections 3.1 through 3.4 and of chapter 1, different performance bottlenecks can be stated:

3.4.1 Diversity of users, applications and performance needs

Researching technical documentation and speaking to the members of team Spatial Statistics showed different types of users within and outside the team Spatial Statistics. This means that the complete geo-ict workload is difficult to assess. Spatial Statistics provides a small number of heavy users, skilled in ArcInfo Workstation and AML and partially in ArcGIS Desktop and Python. Parallel to these users, their colleagues at the department of Process Development and Methodology, use R Spatial for statistical analysis on spatial data. There is no real exchange in knowledge between Spatial Statistics and Process Development and Methodology, although this would be very valuable for both sides: Process Development and Methodology can share their experiences with spatio-temporal data and the application of R Spatial and both teams could exchange lessons learned on functionality and performance of the used software. The light users of Statistics Netherlands apply different software via the virtual desktop, such as ArcView, ArcGIS Desktop and PX Map. This situation makes is difficult to support and optimize the use of spatial data from the point-of-view of the users and the IT department.

[5] Cloud facility of Amazon.com that provides virtual computing services

3.4.2 User experiences with ArcGIS Desktop

A number of users has experienced instabilities with the application, for example crashing of the application after running a tool (no specific tool). The role of the user profile is not really clear: making a new user folder on %appdata% helped. For some tools the desktop application showed a much slower execution time for some users, e.g. the (administrative) JOIN, but also UNION. Al lot of methods to optimize performance and forestall instabilities have been applied: Dataset partition (on attribute or record level), local processing and maintaining the AML /Workstation tools, but there is no sustainable strategy available.

3.4.3 Lack of means of IT department and Spatial Statistics for application management

There are only 2 staff members of the IT department who are doing tasks such as licensing, troubleshooting for the GIS users at Spatial Statistics or even Statistics Netherlands. At the same time, the production processes and available resources at the spatial team leave very few time to fundamentally think about non-functional requirements such as performance, set up SLA's (Service Level Agreements) and look for means to fulfil the non-functional requirements together with the IT department.

The following chapter will present a number of performance factors and analyse their suitability for the benchmark.

Chapter 4: Performance Factors

In the previous chapter we have established a view on organizational and technical bottlenecks of geoprocessing performance by analysing the workload and resources of team Spatial Statistics and Statistics Netherlands. The goal of this chapter is to establish a methodological basis which can be used to develop a benchmark that can provide insight into well performing alternatives within ArcGIS desktop compared to ArcInfo Workstation.

The definition of performance optimization is related to the definition of performance. A viable definition is "the processes of making something, especially a computer system, work as effectively as possible" (Dictionary). Figure 19 shows an interesting schema that provides insight into what Oracle considers factors of influence on database performance and the cost that is implied with the adjustment of performance factors. In the figure, a factor with the highest impact on performance also means that the implementation will come at the highest cost.

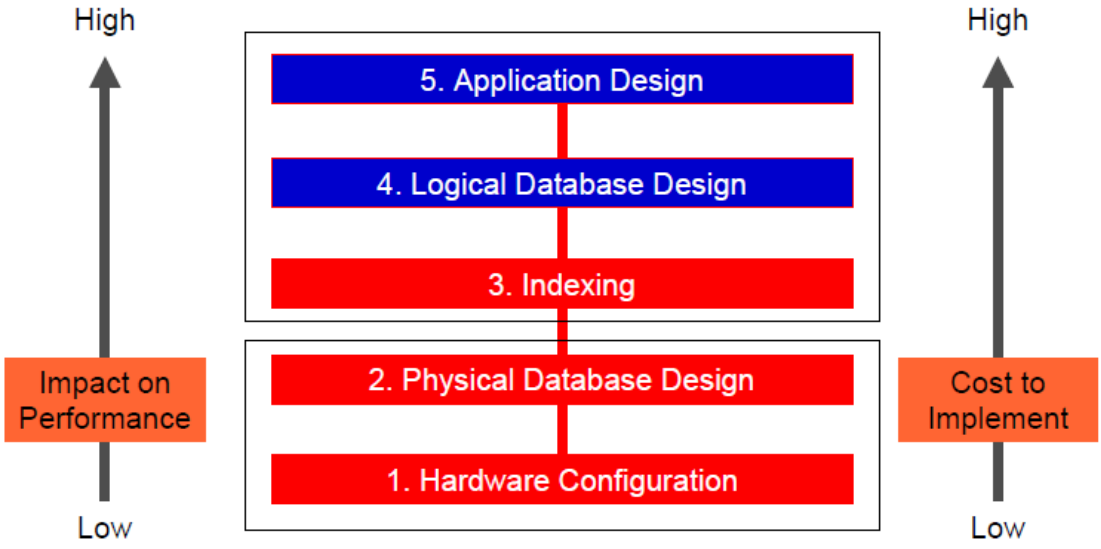


Figure 19: Impact and cost of performance improvement implementation (Godfrind, 2008)

The figure does not state whether the impact is short- or long-term: Changing the hardware configuration and altering the physical specifications of the database could have a high impact on performance, however, concerning long-term performance and scalability, it is more sustainable to look at the application and logical database design. But that would imply to review business goals and the current information architecture. For benchmark design, the logical database design and application design is most difficult and more "costly" to apply.

4.1 Amdahl's Law

To be able to assume the probable optimization that can be achieved with a factor, some basic principles in performance optimization have to be taken into account. A very important principle is the so-called "Amdahl's Law" (Hennessy & Patterson, 2007): This principle is trying to quantify the role of a certain performance improvement for the overall performance of an operation. Speed-up is calculated as a ratio of the execution time of the "slower" version by the execution time of the "faster" version (Hennessy & Patterson, 2007):

$$\text{Speedup} = \frac{\text{Execution time for entire task without using the enhancement}}{\text{Execution time for entire task with enhancement}}$$

To calculate the improvement of an enhancement (a certain factor such as parallel processing or an index), one needs to assume the part of the total that will be affected by the enhancement: if 20 % of the operation can run 10 times faster the ratio would be: 0.2/10. The rest of the operation is 80%, or 0.8 that is unaffected. The unaffected part and the affected part are added up and are used to be divided over 1:1/(0.8 + (0.2/10) = 1,219. This means that 1,219 is the speedup rate of the performance factor (or enhancement). An important lesson from this model is that big overall performance improvements can only be achieved if they affect a large part of the operation.

4.2 Resource configuration

The available resources at Statistics Netherlands combine CPU, RAM, and disk specifications. ESRI provides recommended specifications for its software products, bundled in Table 7: Specifications ARCGIS, ArcInfo Workstation and ArcGIS Pro. The table shows that the hardware demand rises with every new upgrade.

	ArcInfo Workstation 10	ArcGIS Desktop 10.1	ArcGIS Pro
Processor	• Intel Pentium 4, Intel Core Duo, or Xeon Processors; SSE2 minimum • 2.2 GHz minimum; Hyper-threading (HHT) or Multi-core recommended	• Intel Pentium 4, Intel Core Duo, or Xeon Processors; SSE2 minimum • 2.2 GHz minimum; Hyper-threading (HHT) or Multi-core recommended	• Intel Pentium 4/Intel core duo/Intel Xeon processor • Minimum hyper thread dual core • Recommended: quad core • Optimal: 2x hyper threaded hexa core
RAM	2 GB minimum	2 GB minimum	• Minimum: 4 GB • Recommended:8 GB • Optimal: 16 GB
OS	Windows XP, Windows Vista, Windows 7 (32 and 64 bit)	Windows XP, Windows Vista, Windows 7, Windows 8 (32 and 64 bit)	Windows 7, 8 (64 bit)
Disk space	• 820 MB • An additional 100 MB is required for all samples	• 2.4 GB • In addition, up to 50 MB of disk space may be needed in the Windows System directory (typically, C:\Windows\System32)	• Minimum : 4 GB • Recommended: 6 GB or higher • 1.50 GB of available disk space on the installation drive

Table 7: Specifications ARCGIS, ArcInfo Workstation and ArcGIS Pro

The available infrastructure, the fat client as well as the big data computer, have been compared to these specifications (3.3.2.4 Fat clients properties). The available resources at SN, concerning RAM, CPU and disk type, seem to be sufficient.

4.3 Dataset size, scalability

Partitioning the datasets can be very effective in performance improvement as it reduces the time that is spent at scanning the tables and backing-up the tables as well as (re) building indexes. It is also often a preparatory step for multiprocessing. Splitting the tables is often used, e.g. a national census dataset could be split into administrative units, which is called horizontal processing, whereas reducing the number of columns per table would be vertical partitioning. Besides the advantages of partitioning one should, however, also consider the costs of pre-processing and post processing of the data. It also requires thorough knowledge of the data involved and the type of geoprocessing operation.

4.3.1 Horizontal

Horizontal partitioning splits the table into smaller units, often called partitions. A so-called partition key, which is a certain attribute or a combination of different attributes, is used as a partitioning criterion. Reading rows from disk is a relatively slow process, therefore reducing the number of rows reduces the I/O time (Alsultanny, 2010), assuming that the data are clustered in such a way (related to their location) that the data can be retrieved easily.

4.3.2 Vertical

Some tables contain many columns with attributes that are not always needed by the users and also slows down the I/O process. By reducing the number of attributes, the average row length is minimized and therefore less I/O activity is needed (Alsultanny, 2010). Vertical partitioning is mostly executed via normalization (redundant columns are removed and stored in a new table which is linked to the original table) or via row splitting which splits the table directly into 2 or more tables. Also in the case of vertical partitioning, it should be considered that extra cost could be involved in pre- and post-processing. Vertical partitioning divides a table into multiple tables that contain fewer columns. The two types of vertical partitioning are normalization and row splitting:

- Normalization is the standard database process of removing redundant columns from a table and storing them in secondary tables that are linked to the primary table
- Row splitting divides the original table y into tables with fewer columns. Each logical row in a split table matches the same logical row in the other tables as identified by a UNIQUE KEY column that is identical in all of the partitioned tables.

4.3.3 Algorithm and scalability

The algorithm of a geoprocessing operation, together with the data structure, plays an important role in performance of these operations. It determines which steps have to be taken to perform a spatial calculation. In many literature examples, for example in Zhao et al. (2012), the algorithm is built by the researchers themselves, thus more transparent. However, the low level underlying algorithms of the different tools programmed by ESRI are still a black box. Comparison with algorithms for the same tools within other software tool could not be included into the scope of the research.

Therefore, only the “behaviour” of such an algorithm can be analysed with the use of a series of increasing datasets with each of the tools. The role of the algorithm is assessed as vital for performance in scientific literature but also in diverse developer’s websites, such as gis.stackexchange. A very useable comment from a user has been provided (Huber, 2011): “By far the most dramatic improvement in long-running GIS processes is made by improving the algorithm. In many, many cases, if your computation is taking noticeably longer than the time required to read all inputs and write all outputs, chances are you are using an inefficient algorithm.” This not applicable for all computations, investigating the actual performance problem is necessary at first.

4.4 Parallel processing

Data partition, as covered in the previous section, is one of the preparatory steps for parallel processing. Not only are the data partitioned horizontally or vertically, (Qian, 1997) e.g. use a spatial partition of the data

(based on quadtree) to subdivide the data and to process them via a computer grid. A number of publications that deal with processing of large spatial datasets do research on the application of parallel processing as an effective method to improve performance. Zhao et al. (2012) have developed an algorithm for environmental models that calculate statistics of raster layers. Parallel processing is implemented in basically 2 different ways:

1) Executing a computation process, divided in sequences or divided in data, on more than one CPU or CPU core at the same time
2) Executing a computation process, divided in sequences or divided in data, on a network of computers (computer grid)

Before trying parallel processing, a number of aspects have to be considered: Firstly, the ICT configuration has to be suitable for parallel processing. Secondly, the cost of parallel processing is also high regarding needed knowledge, pre-processing, post processing and resources. During parallel processing, the processing steps are processed via more than 1 CPU or computer nodes within a computing cluster, as has been researched in various studies such as e.g. (Abdelguerfi et al., 2005). Using too many parallel units can also lead to more system overhead.

Since ArcGIS 10.1, some geoprocessing modes have been introduced that should help with processing large amounts of data. Firstly, 64-bit processing or background processing has been enabled, providing the only possibility to use 64-bit advantages, since ArcGIS is designed in 32-bit. Secondly, the parallel processing environment parameter has been introduced that is enabled for a number of tools. According to ESRI, multiprocessing is not recommendable for overlay tools because of the Topo Engine, which is the subsystem of ArcGIS that determines the topological relationships of the features claims ca. 60% of the estimated available RAM which is shared by the CPU cores (Pardy & Hartling, 2013). This will clearly exclude the UNION and INTERSECT, being overlay tools. The DISSOLVE and the NEAR would be more suitable. The user can assign a percentage to the parallel processing environment, which is related to the number of cores. 100% means, that all (4) cores will be used. It is recommended by ESRI to use a higher than 100% (more processes than cores), if the processes are I\O bound (ESRI, 2014b).

4.5 Compression and compaction

4.5.1 Compression

Compression of databases is often used to save storage space and network bandwidth (Alsultanny, 2010). ESRI has implemented a compression technique for vector file geodatabase feature classes and tables which makes the data read-only. However, it is not necessary to decompress the dataset before access, because the data is compressed in a direct-access format. During some processes it could provide performance improvement but not all, even slightly less performance in some cases (ESRI, 2013b). Two compression types can be implemented: lossless or non-lossless (lossy). Obviously, in the first option no information is lost, including the preservation of floating point values, in the second option; the data can lose precision, as floating values can be altered. Therefore, this will not be a viable option for Statistics Netherlands, although the compression rate will be higher (lossy compression stands for up to 20% compression). The compression ratio also depends on the type of input data: First, the number of vertices of the features is important. Points and lines with few vertices can compress more than a polygon dataset or a line dataset with lines that contain many vertices. Secondly, attributes field types can be influential on compression ratio: Text, integer and data fields lead to better compression than double or floating field types (ESRI, 2013b).

4.5.2 Compaction

File geodatabases are file based, stored as binary files within a workspace on a disk drive. Editing these data results in fragmentation of the data, this decreases the overall performance. Compacting technique is used to

rearrange and clean up the files within a file (ESRI, 2013a) geodatabase that has been updated frequently. This will also reduce the time used to scan the tables which will improve performance. ESRI advises to perform compaction monthly for file (or personal) geodatabases that are edited frequently and always after a large scale change. The size can be reduced significantly after compaction: about 50%. Eventually disk defragmentation should be performed regularly, for overall disk clean –up (ESRI, 2013a). Disk defragmentation for the local drive is probably not that necessary for the Spatial Statistics users because the data are only stored there temporarily to perform operations.

4.6 Spatial Access Methods

Spatial access methods can be described as a set of different methods that can ensure efficient storage, retrieval and display of 2 or 3 D spatial data, as well as geoprocessing. The two most important aspects of spatial access methods are spatial indexing and spatial clustering. Whereas spatial indexing creates an index file to quickly look-up data and locate them on disk, clustering stores spatial data on disk directly related to their spatial proximity which is often based on methods like space-filling curves. Spatial access methods are often the basis for several geoprocessing algorithms, e.g. spatial selections or overlay operations like UNION or INTERSECT. Spatial indexing is applied at Statistics Netherlands using the default spatial grid index settings for feature classes. Clustering, however has not been applied as the possibilities are limited in ArcGIS Desktop. The clustering techniques like space-filling curves are applied partially on table level (see 4.6.2 Clustering/Sorting).

4.6.1 Spatial indexing

There are different types of spatial indexing, which one is applied varies per GIS software. Other types are R-tree, KD-tree or Field Tree. Spatial indexes are often subdivided into space-driven or data-driven indexes: Space-driven indexes like the quadtree (also used in ArcGIS for loading maps and overlay tools with large datasets) or grid are used to index the MBRs (minimum bounding rectangles) of objects and not the actual geometries. The space is subdivided until the number of rectangles overlapping each quadrant is less than the disk page capacity. Each leaf is associated with a disk page that stores the entries. A rectangle appears in all leaf quadrants that it overlaps. Data-driven indexes like the R-tree are based on the subdivision of the set of objects not the embedding space. The subdivision adapts to the object distribution in the space. Spatial database systems such as PostGIS and Oracle Spatial use the R-Tree. In some cases, the R-tree has a performance advantage to the grid index (Matty, 2012): It is suitable for more dimensions (3D, 4D), uses less storage, is easier to maintain and faster for operations like the Nearest Neighbour. As the data-driven indexes are not implemented in ArcGIS Desktop, this chapter will not elaborate further on these access methods for the benchmark. It is, however, important to consider this: choosing for certain software means choosing also for spatial access methods that are inherent to the software.

4.6.1.1 Grid Index

This spatial index is built by applying a grid to the data in the spatial column. It is possible in ArcGIS to assign 3 grids on different levels with a different resolution (cell size). Dependent on the size of the features, one, two or all three levels can be applied. Figure 20 shows this process: By applying the low-level to the map with polygon nr. 102, the polygon is entered 7 times to the spatial index, in the next level, only 2 times.

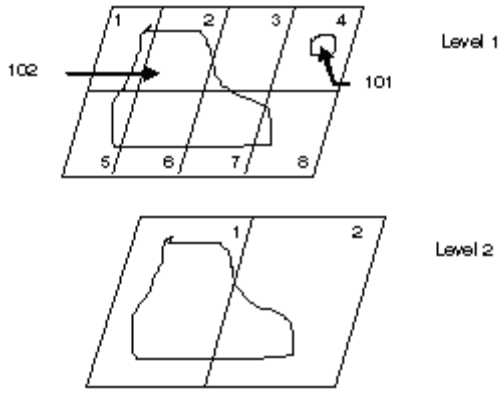


Figure 20: Grid index (ESRI, 2014c)

The selection of the grid size has to be considered very carefully: a size which is too small will result in overhead of index table entries. On the other hand, a size which is too large will retrieve too many features (Harley & Gellerstedt).The grid index has been assessed as inflexible by Oosterom (1999), as it does not cope well with very irregular distribution of data as well as data with polygons that show a lot of variation in size. In ArcGIS, the spatial index is computed and maintained automatically after creation of a feature class and after updating. Only for enterprise geodatabases or databases that use the ST_Geometry storage type in Oracle or Geometry storage in SQL Server, the spatial index can be modified. For feature classes in file geodatabases, which are used at Spatial Statistics, the spatial index can only be deleted and recreated (ESRI, 2013g).

Compressed datasets use a different type of indexing; this is only the case in the compressed mode. After transforming to uncompressed, e.g., the “normal” index is re-established. Manual updating is only advised by ESRI after updating a dataset with a large number of features that differ significantly in size from the other features (ESRI, 2013g). It is possible to enter own assigned grid sizes, although the default grid size is supposed to calculate the optimal grid size index.

For setting the spatial index grid size in ArcGIS, several guidelines have been provided by ESRI (ESRI, 2014c), although not all can be honoured because they are largely aimed at database users (with ArcSDE):

- one grid level is usually enough
- one level with large grid sizes is usually enough for point datasets
- try first one level with a grid size three times as large as the average feature extent size (if the application window is unknown, which is the case in our scenario)

4.6.1.2 Quad Tree

This is a method that is applied with overlay tools that use large datasets, which would not fit into the available memory. At first, it starts with a bounding box that covers the complete dataset. The features are read before sending them to the Topology Engine. Then, the curves are densified and flagged for recreation, followed by an allocation of a large part of memory for this process. Then the second subdivision starts with division of the bounding box into 4 quadrants. All features are read within each quadrant and processed, where necessary, more quadrants will be subdivided until the number of rectangles that overlap each quadrant is less than the disk page memory This is also called subdivision logic. A rectangle can be stored in more than 1 quadrant. Esri uses the quad tree algorithm to perform tiling, a process that divides overlay geoprocessing into smaller portions to make sure that the data fit in the memory. This method is not suitable for parallel processing if features overlap a quadrant.

Figure 21: Tiled overlay processing (Pardy & Hartling, 2013)

The tools that apply this method are e.g.: UNION, CLIP, INTERSECT, IDENTITY, UPDATE, DISSOLVE, FEATURE TO LINE. This approach will not work with large features containing a lot of vertices (many millions of vertices) or if another process is running (ESRI, 2012).

4.6.2 Clustering/Sorting

Clustering is the process of grouping spatial data in such a way that spatially related features are stored in the same memory which leads to faster retrieval of those data and less need to fetch the spatial objects from different, scattered memory locations. Space filling curves have been discovered as a method to assign 2D or 3D data to a 1D memory space (Oosterom, 1999) and (Chen & Chang, 2005). They are organized in tiles and within those tiles follow a certain order. An important characteristic of a space filling curve is that it passes through every point in a certain space once without crossing itself (Chen & Chang, 2005). The most in literature mentioned space filling curves are the Peano and the Hilbert curve, because they show the least number of jumps. The Hilbert curve is more advantageous in that regard because the Peano curve shows a large jump from the node in the left upper quadrant towards the node in the lower right quadrant.

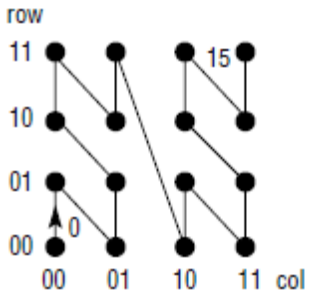


Figure 22: Row prime –right (Oosterom, 1999)

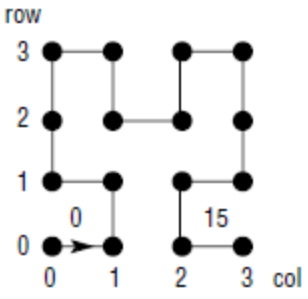


Figure 23: Hilbert Curve – middle (Oosterom, 1999)

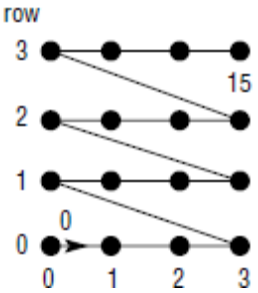


Figure 24: Peano Curve - left(Oosterom, 1999)

A table that is clustered spatially can reduce I/O cost and improve the cost for using the spatial index. If the spatial index has still to retrieve the table rows and then fetch them from different memory spaces, a full table scan can perform better than using the index (ESRI, 2007).
ArcGIS does not employ a high variety in the use of space filling curves; they are implemented in the sorting tool, providing the possibility for row prime space filling curves (UL, UR, LL, and LR, depending on the starting point of the curve) and the Peano curve.

The implementation of those methods varies a lot: Oosterom (1999) states that not many access methods have been implemented for a lot of spatial database systems. Saalfeld (1998) underlines this statement by arguing that many of the existing techniques for analysing spatial data are constrained by the limitations of hardware

and software. It is important to note that these publications are older already: hardware possibilities and software solutions for geoprocessing have certainly improved.

4.7 Data format

The data format of spatial features is often inherent to the GIS software that is implemented, although open standards have been developed over the years. This makes it very difficult to compare performance of operations in different GIS software. Next to the open standard formats like GML, JSON or KML, a number of proprietary formats exists from vendors like e.g. ESRI, MapInfo or Geomedia. As the focus will be on ESRI software, the difference between the coverage and the file geodatabase format (as these are the currently used data formats at Statistics Netherlands) will be analysed and related to possible influence on performance. Both formats are file based, but different in handling topology as well as attribute data.

The definition of polygons as arcs, which represent an ordered set of edges made it easier to store large polygons. The storage is efficient as well: an edge could be used for more than one polygon.
On the other hand, the "on-the-fly" assembly of edges to polygons could also slow down some geoprocessing operations, involving e.g. multi-part polygons (regions) and multi-part line features (routes). Moreover, editing of the coverage implies rebuilding of topology (Esri, 2004). Storage of attribute data in INFO tables is also efficient for "administrative" processing. Only the INFO tables have to be read and written to memory.

Topological data formats have advantages and disadvantages regarding topology: In general, explicit topology is leading to slower performance in rendering and displaying spatial data because the vertices have to be read from the different tables. Moreover, maintenance of the topology after updating a file can be time consuming. A large advantage of explicit topology is the execution of certain geoprocessing overlay operations, because the explicit topology reduces storage and saves substantial computation time (Theobald, 2001). This is also the case for an aggregation tool like the DISSOLVE: The principles of contiguity and area definition storage of adjacency enable the application to search (indexing and sorting necessary), retrieve and delete the common arcs of a group of polygons. Because of its topological structure and its legacy status, there are also certain limits. These are some examples of limitations (ESRI, 2010b):

- Max. 500 coordinates per arc (after the 500th coordinate, a new arc is created)
- Max 360 arcs per node
- Max. 10000 INFO tables per workspace
- Max. 4,096 bytes or characters per records within a feature attribute table cannot exceed
- Processing limitations can occur during cleaning and building the topology and overlay tools:

4.8 Application Design

The factor of application design is the least transparent, because different aspects within the development of an application can influence performance: Algorithm, interoperability with operating systems, interoperability with big data storage applications, 64bit design, use of hardware resources and implementation of spatial access methods (related to algorithm).

Some methods that are implemented are known, for example the spatial access methods that ArcGIS applies (Grid index, row prime and Peano sorting, quad tree partition of data during overlay processing), but the low level programming architecture is largely unknown, which makes it impossible to use application design itself as a factor within the benchmark. In some examples like (Sorokine et al., 2012), disappointing results of a certain algorithm in ArcGIS Desktop 9.x (in comparison to ArcView) led to low level design of the algorithm in C++ and Java, leading to satisfying results for the low level application.

The application design is more of a framework for the implementation of the factors (which performance factors can be implemented in ArcGIS), whereas the results of the benchmark will hopefully provide valuable insight whether the application design provides enough possibilities to keep performance at an acceptable

level. Within the area of low level application, a number of studies are also conducted to analyse impact of implementing geoprocessing algorithms directly within programming languages such as C++ or Java which results in significant performance improvement (Sorokine et al., 2012).

4.9 Network

In chapter 1 it became clear that the heavy calculations are executed locally on fat clients: the executable is on the local drive as well as the data that have been copied to that location. What remains on the network is the script, the user profile and the license server. It is not expected to have a significant impact, although it is still advisable to test a script and one tool to quantify the role of the network within performance. The fact remains that the fat clients have to use a real line via the WAN to the data centre, which is could lead to delay in performance.

4.10 Use of Workspace

Within the ESRI software, the term "workspace" is used to indicate a "container for geographic data" (ESRI GIS Dictionary 2015), which could be a file geodatabase, a feature dataset, a folder or an ArcInfo workspace. It has been indicated by ESRI that the storage of input dataset and output datasets in different workspaces (directories) could have impact on geoprocessing performance. This is not documented, but it could be a "noise" factor that has to be excluded by comparing performance of geoprocessing using input and output data with different workspaces with the use of the same workspace.

4.11 Discussion

A lot of information has been collected on benchmarking methods, performance factors and the workload and resources of Spatial Statistics. Not all information could be retrieved: For research question 1, some information was not present, e.g. hardware resource usage: these metrics have not been collected by the staff in their working process. More important, the current business objectives of Spatial Statistics are clear, but not the objectives over a longer time, e.g. 5 – 10 years. Therefore, suitability of ArcGIS Desktop or ArcGIS Pro infrastructure can be evaluated for current geoprocessing processes only (or rather – a part of those processes).

The actual benchmark has to be built on the result of chapter 2, 3 and 4. However, not all factors could be covered in depth, because the range of the factors is very broad. Often, one factor alone, such as spatial indexing, even one type of index, is covered extensively in related work such as Arge et al. (2002), Oosterom (1999) and Mellor-Crummey (2001). Contrary to part of the related work researched for this benchmark, the factors could not be covered in depth in this report, as in-depth research per factor would result in different separate research projects. Moreover, many examples are not wholly applicable for real life organizations because of the restraints that are inherent within an organizations infrastructure, policy and capacity. For example, there is a modest number of spatial database benchmark or comparisons: Tijssen (2012), Ray (2011), Paton et al.(2000), Batcheller et al.(2007) and Matty (2012). However, except for Batcheller et al., none of these benchmarks has included ESRI products such as ArcInfo Workstation and ArcGIS desktop. The publication of Tijssen et al included ArcSDE middleware, but the DBMS at the basis is Oracle. The study of Matty (2012) compares Oracle spatial with Postgres Spatial. Additionally, the related work listed in this section covers spatial DBMS products, which is also not applicable for many organizations since they use ArcGIS Desktop products mostly and work with file geodatabases. Therefore, the restriction of the scope to ArcInfo Workstation, ArcGIS Desktop and ArcGIS Pro also excludes a number of factors. Without a full spatial DBMS, for example, such as Oracle or Informix, there are no possibilities to fine-tune database performance, e.g. the adaption of resources such as page size or modifying the spatial index.

The file based ArcGIS Desktop provides other possibilities to improve performance which have not been researched in a scientific project before: compacting and compressing of input data, adjusting the spatial index, 64bit background processing or the new desktop product ArcGIS Pro. Hardware and software related factors are strongly interrelated: The review of performance factors in previous sections has shown that every factor

can help to use the available hardware resources in a more efficient way. The question is to find out which factor or combination of factors that is implementable within ArcGIS (desktop) will be equal to the current performance in ArcInfo Workstation or even outperform it. What is yet unclear is the factor that makes certain operations in ArcInfo Workstation faster than in desktop: the data model (coverages) with its explicit topology or the application architecture which possibly uses the system resources differently from ArcGIS Desktop. This is still a “black box”, although some assumptions can be made regarding the role of explicitly stored topology.

The first priority therefore is to compare scalability between ArcInfo Workstation and Desktop: The data model as well as the efficiency (strongly related with the data model) of the algorithm. The efficiency of the algorithm could be made visible with the help of a controlled, synthetic dataset in different sizes (number of records). The synthetic data would have to be scaled considerably to provide enough information (e.g. from 10.000 records up to 1.000.0000 records).

The second priority is to exclude “noise” that could disturb performance: Network interference, although considered a small factor and location of the input data in different workspaces. The network is expected to have the least impact, because it is used for reading the script, writing the logs, and the license server only. Maybe the roaming profile (see chapter 3) on the C-drive can be of influence. It won’t be necessary to test all tools with the network factor.

Finally, a number of optimization factors can be evaluated in ArcGIS 10.1 on the fat client to assess how much optimization is possible with the available resources of Spatial Statistics: for example compression and compaction, indexing and sorting and spatial indexing. For indexing, it is not known how the default index is calculated. The default index can be compared with input datasets without a spatial index, or a new index with assigned grid levels. Partitioning of the data as well as parallel processing is not suitable for all workload scenarios. Therefore, the partition has less priority than compression/compacting, indexing and sorting.

Referring to the schema of Godfrind (2008), the hardware implementation is the easiest implementation, however with least (long term) impact. For the situation at Statistics Netherlands, there is a certain degree of bias, because the hardware implementation is strongly related with the software version. The other way round is true as well: ArcGIS Pro is very promising as a true 64 bit GIS, but only installed on a computer with much higher specifications than the “regular” fat client.

This benchmark should provide some guidance for staff outside the academic ICT research facilities to evaluate performance within a non- or less configurable environment (such as the fat clients) and with limited ICT support and limited optimization possibilities within the software.

Chapter 5: Benchmark Development for Statistics Netherlands

5.1 Method

It has been clear at an early stage of this research project that the use of a macro workload/benchmark, which would test the performance of a complete production process of one of the products of the spatial team, would be very costly in time, transparency (regarding the identification of possible performance bottlenecks) and repeatability. A number of products are in the stadium of redesign, including the analysis method. Therefore, the "micro benchmark" approach of testing single tools has been chosen for the tools that have been indicated by Spatial Statistics and that would render interesting results during base line tests.

The main goal of the benchmark is to provide information on performance optimization of the current workload within ArcGIS desktop and ArcGIS Pro. The benchmark has to yield reliable results but also reflect the current workload. The reliability is reflected in the following measures:

- Create a micro benchmark containing single geoprocessing tools
- Combine 1 factor and 1 tool only (except for the resource related factors)
- Validation of the results: performance measurements and output data

The workload is reflected in:

- The use of frequently used real datasets of different size
- The use of large synthetic datasets

The list of factors is already very extensive, but only a selection of those factors is covered extensively in literature, even a smaller selection is applicable in ArcGIS Desktop and ArcGIS Pro.

5.2 Performance factors

The first step of the benchmark is the baseline test, first in ArcInfo Workstation to establish the performance requirements (in execution time), secondly in ArcGIS Desktop 10.1, to establish the performance, execution time and CPU usage in default mode. Consequently, the scalability and stability of the algorithm of the chosen geoprocessing tools UNION, INTERSECT, DISSOLVE and NEAR will be tested with synthetic data in ArcGIS 10.1 on the fat client (operating system Windows 7).

As a next step, possible "noise" as a result of network interaction has to be excluded. Therefore, only one or two tools in combination with a large dataset, are sufficient to detect possible influence. The following step will be an "exclusional" factor as well: Does storage of input datasets in one file geodatabase (workspace) differ from storage in separate workspaces? To exclude that impact, only 1 or 2 scenario's are sufficient as well.

The following three factors, spatial index, sorting and compacting/compressing are optimization factors: Can they optmize performance of the tools in ArcGIS 10.1? Therefore, the real dataset will be used,but not for every tool: The sort will be most useful for DISSOLVE or NEAR, whereas the use of a spatial index could be of importance for all geoprocessing tools.

The last two factors will be executed on the big data computer: one scenario to compare the default ArcGIS 10.1 fat client performance with the big data computer (ArcGIS 10.2.2) performance, and one to compare ArcGIS Pro, a 64 bit application, with the results of the previous step (ArcGIS 10.2.2)

These steps are visualised in Figure 25: Performance factors. Each color corresponds with a benchmark step:

Green	Baseline and scalability
Orange	Exclusion of network and workspace influence
Purple	Optimization of performance in ArcGIS 10.1
Blue	Change of hardware/software configuration

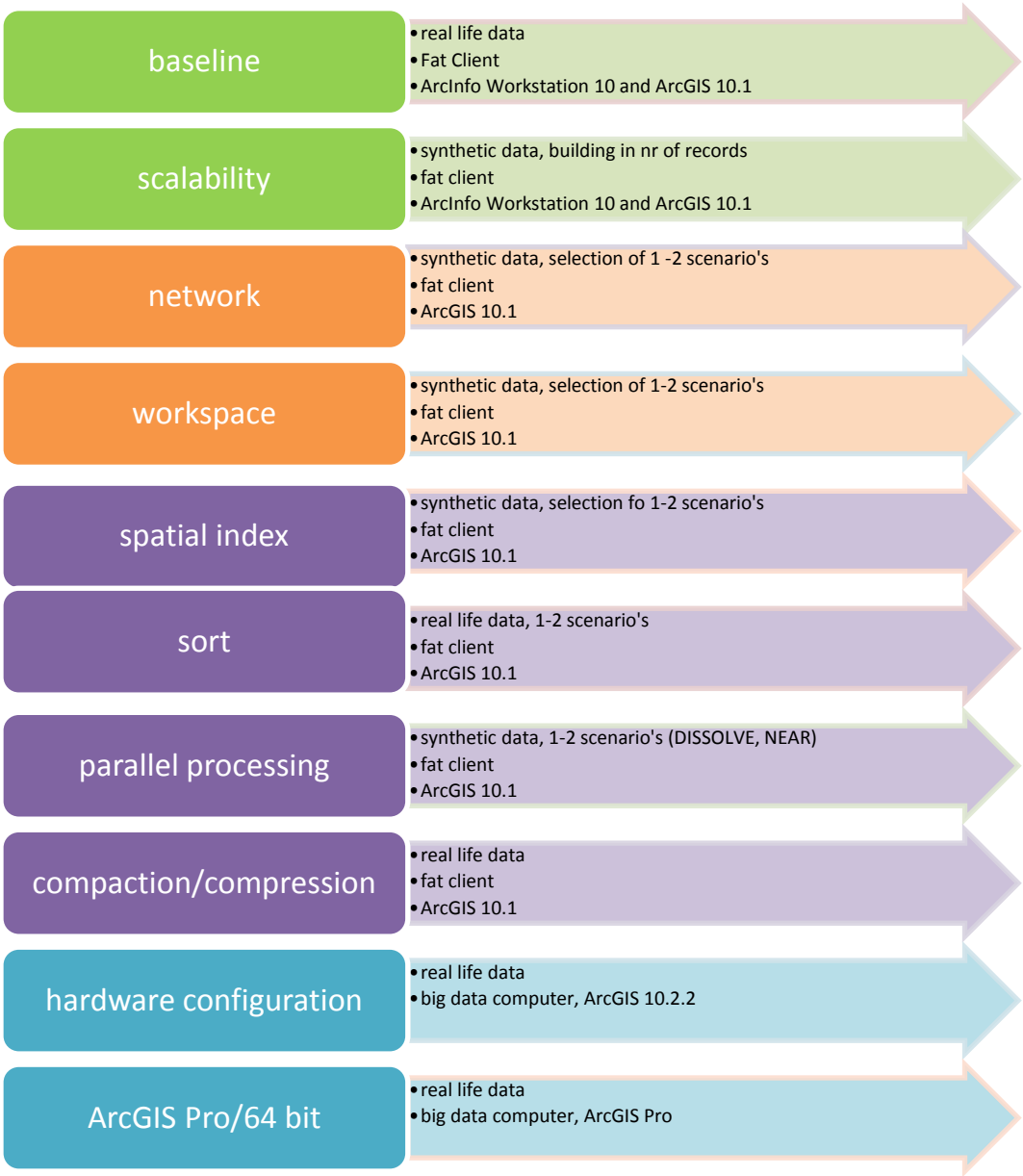


Figure 25: Performance factors

5.3 Analysis and Validation of the results

5.3.1 Analysis of execution time and resource usage

The execution time is recorded 5 times per operation to get an impression of the consistency of the results. Which statistical value should be used of the 5 time results, depends on the regularity. If big outliers are recorded, the mean value is not a good option, but rather the median. In many cases, the first run could result in a slower execution time because the data have not been loaded into memory. Therefore, the difference between first and subsequent value(s) should be observed closely in the log file.

5.3.2 Validation of the output

As established earlier, not only the performance (execution time, resource usage) is of importance in this benchmark, but also the quality of the results. The same query has to result in the same feature class or table. To check the quality of the results, the 5 output datasets (5 runs per tool) will be checked on the number of records per output dataset. Ideally, the output should also be checked on:

- Content of attributes – Does a query on one or more attributes yield the same results?
- Check on geometry: compare the area of the largest and smallest polygon between the datasets, check the dataset on overlapping features, slivers and overshoots.
- Check consistency of statistics of geometrical fields, e.g. area, length.

The number of test runs will be high, therefore structured checks on these points will be very time consuming. Therefore, the check on the number of records per output is a priority and will be carried out first.

5.3.3 Test data and metrics

It is important to distinguish between data and metrics. The raw data are the result of data collection during the logging process such as e.g. the number of seconds of execution time and CPU time.

A couple of aspects need to be considered for the evaluation of the logging data:

- Not every tool supports the implementation of a factor and a factor can be applied in different variations (size datasets, application of tool for one or two input datasets and other tool parameters)
- Although the logging results are quite regular, rounding and calculating the medium value will result in less precision of the results. Therefore, the resulting values should be studied carefully, by checking the original logging data.
- Foremost, stability is most important, because SN has to deliver reliable statistical products.
- Execution time of the optimization factor related to the execution time of ArcInfo Workstation is important because of production lead times that have to stay the same, but not all scenarios could be tested in ArcInfo Workstation.
- The resource usage has not been included in the score card because there is not much known about threshold values for CPU time and page faults for spatial queries. Moreover, resource usage could not be measured for ArcInfo Workstation.

5.4 Setting up the test bed

The benchmark has to be as close as possible to the real world situation workload, but also provide insight into the processing mechanisms and scalability of the tool itself, without the distorting effect of specific datasets and their irregularities. Based on the results of chapter 2, 3 and 4, a micro benchmark with the tools DISSOLVE, UNION, NEAR, INTERSECT and JOIN FIELD will be set up, using real and synthetic data. Therefore, the (geoprocessing) tools (DISSOLVE, UNION, NEAR, INTERSECT) will be tested first with synthetic data, in this case grid polygons as well as their label points. An example of the synthetic input datasets is shown in Figure 26. For most of the input data, the data preparation process has been documented in model builder. The model will be delivered with the digital version of this document as a .tbx file, but is also shown in **Fout! Verwijzingsbron niet gevonden.**.

The synthetic datasets are grid polygons and their label points, created with the ArcGIS tool “CREATE FISHNET”. For the DISSOLVE input, fishnet grid datasets of 25 m^2 grid size and 100 m^2 are combined with a UNION. This dataset contains 150.800.000 records. The DISSOLVE field is the ID of the 100 m^2. Out of these UNION results, smaller datasets can be selected, ranging from 10.000 records up to 50.000.000 records. 50.000.000 records have been chosen as a maximum due to the long preparation time. Every fishnet grid also creates a dataset of label points. For the NEAR, a selection of the 25 m^2 label points, ranging from 10000 up to 500.000 records has

been used with a 100 m^2 fishnet grid line dataset, containing 9.412.000 records. For the INTERSECT and the UNION tool input datasets of100.000, 500.000, 1.000.000, 5.000.000, 10.000.000 and 50.000.000 will be used in combination with a grid dataset of constant records size (original dataset), in order to alter only one parameter instead of two. The synthetic input data will have a default spatial index, no projection.

For the baseline tests, the real life datasets are already prepared for use by Spatial Statistics, for example, the NWB or the BAG tables. Other datasets, such as the neighbourhood (WB) or the land use (BBG) datasets, can be used instantly. However, further data preparation is needed to create different input datasets per factor. To apply compressing and compacting, for example datasets that are compressed, compacted, or both have to be created. For data partitioning and parallel processing, smaller data sets have to be created with FEATURE CLASS TO FEATURE CLASS.

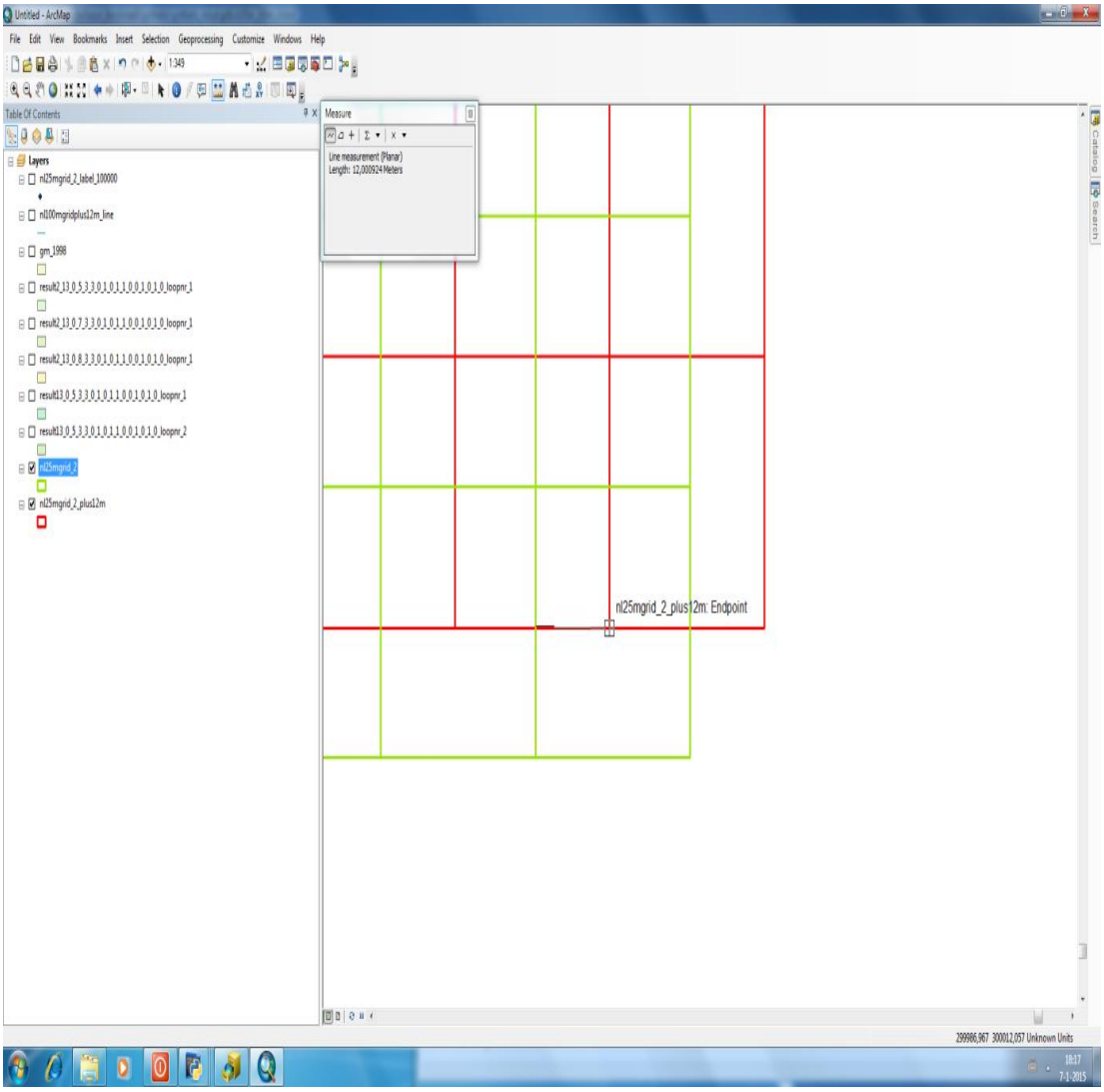


Figure 26: Example synthetic data - input datasets UNION and INTERSECT

These tools will be tested isolated, in a similar fashion as the micro workload, chosen in Jackpine (Ray et al 2011). Combined with the datasets the workload scenario's shown in tables 8 and 9 will be tested (table 9 only for resource related factors):

Tool	Dataset real (name and number of records)	Dataset synthetic[6]	Geoprocessing scenario (real and synthetic)
DISSOLVE	BBG 2008: (233.153 rec.)	• 100.000 rec. • 500.000 rec. • 600.000 rec. • 700.000 rec. • 800.000 rec. • 900.000 rec. • 1.000.000 rec. • 5.000.000 rec. • 1.000.0000 rec. • 50.000.000 rec.	• Real: Dissolve BBG 2008 (bbg2008) boundaries based on attribute value of field "bg2008a" • Synthetic: Dissolve nl25m_100m_Union selection on field "FID_nl100mgrid"
UNION	• District/Neighbourhood boundaries 2012 (12.882 rec.) • District/Neighbourhood boundaries 2008 (12.570 rec.)	• Input 1: fishnet grid of 150.000.000 rec. • Input 2: fishnet grids with different bounding box): • 100.000 rec. • 500.000 rec. • 1.000.000 rec. • 5000000 rec. • 10.000.000 rec. • 50.000.000 rec.	Create union dataset of : • (real)District/Neighbourhood boundaries 2012 with District/Neighbourhood boundaries 2008 • (synthetic) 2 sets of fishnet grid (same number of records, different bounding box)
INTERSECT	• NWB (2.826.593 rec.), • Forest land use (25.157 rec.)	• Input 1: fishnet grid of 150.000.000 Rec. • Input 2: fishnet grids with different bounding box): • 100.000 rec. • 500.000 rec. • 1.000.000 rec. • 5000000 rec. • 10.000.000 rec. • 50.000.000 rec.	Calculate intersections of: • (real) NWB lines and as "forest" classified BBG polygons • (synthetic) 2 sets of fishnet grid (same number of records, different bounding box)
NEAR	• NWB (2.826.593 rec.) • Intersections of forest areas and NWB (126.331rec.)	• Input 1 : fishnet grid converted to polylines of 9.412.000 records • Input 2: Label points of fishnet grids: • 10.000 rec • 25.000 rec. • 50.000 rec. • 500.000 rec	• (Synthetic): calculate within 5000 m the nearest projection of a 100 m^2 fishnet grid line dataset from a label point dataset, varying in nr. Of records. • (Real): calculate projection points on NWB within 5000m from the intersection points of Top10NL with the land use file (as "forest", classified polygons. The value for forest in BBG is "60". The Filename of theses intersection points is "intersect_p")

Table 8: Workload scenario geoprocessing

[6]Fishnet grid (created by UNION of 25 m^2 and 100 m^2 grid size fishnet grids):

Tool	Dataset	Geoprocessing scenario
JOIN FIELD (optional)	• House numbers (8.668.742 rec.) • combined dataset of residential objects, pitches and anchorages (8.443.296)	JOIN FIELD based on ID
FREQUENCY (optional)	• House numbers (8.668.742 rec.)	Calculate frequency of field "straat" and store the results in an output table
SUMMARY STATISTICS (optional)	• House numbers without values "NULL" (8.443.273 rec.)	Calculate the means of fields "POINT_X"and "POINT_Y" per PC6[7] to determine PC6 centroid and store the results in an output table

Table 9: Workload scenario administrative processing

As part of the testing environment, the following software is needed:

- ArcGIS 10.1 (and 10.2)
- ArcGIS Pro beta (Pre-release version)
- ArcInfo Workstation (including technical assistance from SN)
- AML (including technical assistance from SN)
- Python 2.7 and 3.1

Python is used to automate and document the processing steps, but also to record performance metrics, such as execution time, CPU use, etc. The Python libraries (including technical assistance from SN) will be given in Table 10:

Library	Goal
Arcpy	ArcGIS 10.x geoprocessing
Datetime	Records date and time
Psutil	Records usage of resources
Sys	System related commands
Traceback	Errorhandling

Table 10: Required python libraries

The script developed for the benchmark will write the results of each scenario to a log database. The scenario's will be marked with a unique ID, built by the different factors as component :

A_B_C_D_E_F_G_H_I_J_K_L_M_N_O_P. For each letter (variable), a domain of numbers has been assigned, shown in Appendix 3: Nomenclature logfiles. For example variable A has a domain of numbers assigned to tools (e.g. 2 for a DISSOLVE, 4 for an INTERSECT). After a number of test runs, the variable of warm/cold run dit not prove to be important for the benchmark measurements. Whereas each ID stands for a speficic scenario, an additional number per run will be added to the ID to distinguish the results per run.

[7] Postal Code areas (on 6 digit level, in Dutch "Postcode 6")

A	Toolname
B	Hardware configuration
C	Population length
D	Population width
E	Index
F	Spatial sort
G	Filet type
H	Compressing/compacting
I	Daytime
J	ArcGIS version
K	Processing
L	Location script
M	Cold/warm run
N	Projection
O	Real/synthetic data
P	Workspace

Table 11: Nomenclature logfiles

5.5 Defining the performance level

During interviews with the GIS staff, the most important demand for a migration to e.g. ArcGIS desktop 10.1 is that execution time of the tools should at least not last longer than currently in ArcInfo Workstation, apart from other non-functional requirements such as stability and a correctness of the output data. Out of the non-functional requirements, the performance (in execution time) had highest priority, although scalability should be regarded as equally important. Table 13 shows an indication of the performance needs, expressed in execution time (seconds). In this table, dataset and size expressed in memory and number of records has been added because they are essential factors for the level of performance:

Tool	Datasets	memory	Number of. Records	Win7/WS10 (sec)
DISSOLVE	(land use map) BBG2008	601 MB	233.153	**32**
INTERSECT	• (enriched road map) nwb_split • (land use file/wood) bos_bbg08	• 3,49 GB = 3573.76 MB •	2.826.593 (NWB), 25157 (BBG_bos)	**518**
NEAR	• NWB_split	• 3,49 GB = 3573.76 MB	2.826.593 (NWB), 126. 331 (intersect_p)	**14.707**
UNION	• NEDWB_08 • NEDWB12	• 18,7 MB • 27,5 MB	12.570 (WB2008), Xxx	**47**

Table 12: Indication performance requirements geoprocessing tools

The heavy users within the spatial team have indicated the current execution time of the tools in ArcInfo Workstation as a very clear requirement for a more up-to-date GIS configuration. Therefore, the table which has been the result of the initial baseline in Workstation, will be shown again to show the minimum performance requirements. The following table shows the performance requirements for a number of administrative processing tools.

Tool	Datasets	Memory	No. records	Win7/WS10 (sec)
admin. Join indexed (JOINT ITEM/JOIN FIELD)	• House numbers (num2012ba) • vsl2012ba.pull	• num2012ba:552 MB • vsl2012ba.pull: 587 MB	(num2012ba) 8668742, (vsl2012ba.pull) 8443296	**320**
FREQUENCY	House numbers (num2012ba)	811 MB	8.668.742	**385**
SUMMARY STATISTICS	House numbers without values "NULL"	811 MB	8.443.273	**102**

Table 13: Indication performance requirements administrative processing

At first, the tools will be tested within the "default" settings, meaning that there will be test runs per tool with the configuration that the staff works with: For example, with the use of file geodatabase as data format, with the (default) spatial index. With a file geodatabase, a spatial index is always created, therefore, testing it without a spatial index is not realistic, although the regularly used ArcInfo Workstation does not use spatial indexes in default, only by using the AML command "build index". Other factors are not used within the daily processing activities. It is difficult to predict performance development in execution time from baseline ArcInfo Workstation towards ArcGIS Desktop 10.1. Some tools such as the DISSOLVE are much faster than in ArcGIS 10.1. On the other hand, the coverages ArcInfo Workstation are limited in a number of properties such as file size. Moreover, the UNION e.g. did not run stable in Workstation on the BBG and had to be scaled down to a smaller file. The execution time per baseline scenario will be recorded, but also CPU time. The scenarios for the real life workload are described in Table 8 and in Table 9. During the scalability scenarios's the synthetic data (creation of the data described in section **Fout! Verwijzingsbron niet gevonden.**) will be used. The four tools DISSOLVE, UNION, INTERSECT and NEAR will be executed with growing input datasets (only one input dataset will grow in number of records). Except for the NEAR, a fairly high number of records has been used. The synthetic workload scenarios are described in Table 8. Most of the tests will be in ArcGIS 10.1 Desktop, only the DISSOLVE will run in ArcInfo Workstation as well, for grid datasets of 100.000, 500.000 and 1.000.000 records, because creation of the other input datasets caused errors. For these errors, no bug reports have been issued because the limits of creating polygons in ArcInfo Workstation are known.

5.6 Scenario's to exclude noise from the network and workspace

5.6.1 Location of the script

In chapter 2 we established that both scripting languages, AML and Python, are interpreted implementations that have to be compiled during execution. As scripts are mostly accessed via the network, performance is expected to improve for a complete script. The script that executes the benchmark will be used for that purpose and execute the DISSOLVE tool on the synthetic data set of 50.000.000 records. It is not known how the script will be executed: will the complete script be read at once, loaded into the local memory and executed locally, or will the script be retrieved and executed line by line. In the last case, significant performance improvement should be measured.

5.6.2 Day/Time Execution

In chapter 1 it has been established that several elements of the geoprocessing process have to use network capacity, such as reading, compiling and executing the script, accessing the user profile and the license server. This factor should be applied to one or two tool-data combinations with a long execution time at the following day/time combinations:

a. Monday/Friday during office hours, e.g. 11 am
b. Monday/Friday/ after 7 pm or weekend

Tool	Dataset	Dataset size in byte	Nr of records
NEAR	NWB, Park Entrances (intersect_p), both including spatial index	498.427.106 NWB, 14.953.331 Intersect_p	2826593 (NWB), 126331 (intersect_p)

Table 14: Scenario network

5.6.3 Input and output in same workspace

Like the previous scenario's, one or two tools with a large input datasets are sufficient. The DISSOLVE (50.000.000 records) and the NEAR (100.000 label points) will be used for this purpose. The impact is expected to be very low.

5.7 ArcGIS 10.1 optimization scenario's

5.7.1 Implementation Sorting

Sorting the data using a space filling curve such as the PEANO curve can help in drawing performance in ArcMap as well as faster routing and directions when compared to un-sorted data. In this benchmark, the PEANO curve sorting algorithm will be applied as a factor to improve performance of the NEAR and the DISSOLVE tool. With the NEAR the sorting will be applied in two variations: sorting of the point input dataset and sorting of both input datasets. The DISSOLVE will also use two variations but in a different way: for the first variation only the spatial sorting will be applied, for the second spatial and administrative sorting will be combined.

a. Default (no sorting)
b. PEANO sorting

Tool	Dataset
Near	NWB and intersection points (without sorting)
	NWB and intersection points (with PEANO sorting, ascending)
Dissolve	BBG2008 (without sorting)
	BBG2008 (with PEANO sorting, ascending)
	BBG2008 (with PEANO sorting and sorting of field "bg2008a", ascending)

Table 15: Sorting workload scenario

5.7.2 Implementation Spatial Indexing

As explained in section 4.6.1 Spatial indexing), ArcGIS Desktop does not provide the possibility to view and adapt the grid levels for the spatial index. It is only possible to delete and recreate the index. Therefore, this approach is intuitive and black box: Does the spatial index matter in this scenario? The effect of the spatial index (default, modified and without) will be tested. The modified version will depend on the spatial distribution and size of the feature. The quadtree subdivision that is applied for overlay tools such as UNION and INTERSECT will be automatically applied for large and cannot be "turned off". Table 16 shows the input data sets.

Tool	Factor variation	Dataset	Number of records input datasets
NEAR	Default spatial index	Synthetic data set with label points of 25 m grid as input features and 100 m grid polyline as NEAR features	• 10.000 • 9.412.000
		Synthetic data set with label points of 25 m grid as input features and 100 m grid polyline as NEAR features	• 25.000 • 9.412.000
		Synthetic data set with label points of 25 m grid as input features and 100 m grid polyline as NEAR features	• 50.000 • 9.412.000
	Adapted spatial index 1000_0_0	Synthetic data set with label points of 25 m grid as input features and 100 m grid polyline as NEAR features	• 10.000 • 9.412.000
		Synthetic data set with 25.000 input points and (label points of 25 m grid) and 100 m grid line	• 25.000 • 9.412.000
		Synthetic data set with 50000 input points and (label points of 25 m grid) and 100 m grid line	• 50.000 • 9.412.000
	Adapted spatial index 25000_0_0	Synthetic data set with 1000 input points and (label points of 25 m grid) and 100 m grid line	• 10.000 • 9.412.000
		Synthetic data set with 25000 input points and (label points of 25 m grid) and 100 m grid line	• 25.000 • 9.412.000
		Synthetic data set with 50000 input points and (label points of 25 m grid) and 100 m grid line	• 50.000 • 9.412.000
	No spatial index	Synthetic data set with 1000 input points and (label points of 25 m grid) and 100 m grid line	• 10.000 • 9.412.000
		Synthetic data set with 25000 input points and (label points of 25 m grid) and 100 m grid line	• 25.000 • 9.412.000
		Synthetic data set with 50000 input points and (label points of 25 m grid) and 100 m grid line	• 50.000 • 9.412.000

Table 16: Scenario spatial index

5.7.3 Parallel Processing

For the evaluation of this factor, only DISSOLVE and NEAR have been chosen, because the environment setting is not suitable for overlay tools such as INTERSECT and UNION.

Tool	Parallel or linear	Percentage core usage	Dataset
DISSOLVE	Linear	-	Synthetic dataset 50.000.000 records
	Parallel	Default – tool decides how processes are distributed	Synthetic dataset 50.000.000 records
	Parallel	50%	Synthetic dataset 50.000.000 records
	Parallel	75%	Synthetic dataset 50.000.000 records
	Parallel	100%	Synthetic dataset 50.000.000 records
Near	Linear (default)	-	synthetic datasets: point dataset of 100.000 records, Grid polyline dataset of 9.412.000 records
	Parallel	Default – tool decides how processes are distributed	synthetic datasets: point dataset of 100.000 records, Grid polyline dataset of 9.412.000 records
	Parallel	50%	Synthetic datasets: point dataset of 100.000 records, Grid polyline dataset of 9.412.000 records
	Parallel	75%	Synthetic datasets: point dataset of 100.000 records, Grid polyline dataset of 9.412.000 records
	Parallel	100%	Synthetic datasets: point dataset of 100.000 records, Grid polyline dataset of 9.412.000 records

Table 17: Scenario parallel processing

5.7.4 Compacting/Compressing implementation

These possibilities will be offered separately and combined. The lossy compression will not be used because Spatial Statistics has to maintain data precision.

a. Compressed (lossless)
b. Compacted
c. Compressed (lossless) and compacted

Compressing will most probably not have an impact on performance, it reduces the file size, but this could only influence the I/O process and not the calculation process. Moreover, it is not known how much time it takes to remove the read only status of the compressed file. If an assumption has to be made, the execution time for compression will remain at the same level as the default settings. It is not expected that compacting has a very important role in improving performance, because the data that are mostly used, are not frequently updated, with the exception of the BBG, which is the result of inserting and editing land use polygons. It also does not always lead to better performance, although it does saves storage. The NEAR will be tested with a compacted file only, as the compression on the roadmap (NWB) resulted in an error.

Tool	Dataset
DISSOLVE	BBG2008 default, not compressed and not compacted
	BBG2008 compressed loss-less
	BBG2008 compacted
	BBG2008 compresses loss-less and compacted
UNION	District/Neighbourhood boundaries 2012 and District/Neighbourhood boundaries 2008 default, not compressed and not compacted
	District/Neighbourhood boundaries 2012 and District/Neighbourhood boundaries 2008 compressed loss-less
	District/Neighbourhood boundaries 2012 and District/Neighbourhood boundaries 2008 compacted
	District/Neighbourhood boundaries 2012 and District/Neighbourhood boundaries 2008 compacted and compressed loss-less
NEAR	NWB and intersections (intersect_p) default, not compressed compacted
	NWB and intersections (intersect_p) compacted

Table 18: Datasets compacting and compressing

5.8 Resource upgrade scenario's

5.8.1 Implementation Hardware and Operating System

The hardware configuration is expected to show some improvement in performance, although it is not substantial according to Godfrind (2008), when compared to application design or database structure. However, compared to the fat client configuration, especially the SDD technology is expected to speed up the analysis process. The factor is hard to determine, data access can be up to 100 times faster. The RAM memory is about 4 times larger, but without 64 bit processing, ArcGIS will probably not be able to use such an amount of memory, although the benchmark datasets will fit in the RAM memory. The clock speed of the CPU is lower, although there are much more cores and more memory assigned to the CPU. Table 19 shows the main properties of the fat clients of the team Spatial Statistics and the big data computer in the innovation laboratory.The real data scenarios will be executed on the big data computer in ArcGIS 10.2. This is not directly comparable with ArcGIS 10.1.

	Fat client	Big data computer
Operating system	Windows 7	Windows 7 professional
GIS software	ArcGIS 10.1 SP1	ArcGIS 10.2
Hardware provider	Fujitsu	Fujitsu Celsius M720
Disk	HDD	SDD (4x256 GB SSD)
Windows performance index	5.2	5.4
System type	64 bit	64 bit
RAM	8 GB (7,85 GB available)	32 GB
Processor type	Intel® Core™ i5-3570 CPU @3.40 GHz 3.40 GHz (4 cores)	Xeon E5-2640 2.5 GHz 15MB Turbo Boost, 16 cores
C drive	Part of the C drive (user and temp folder) reconnects to the datacentre	C- and E-drive (both local hard disk drives)

Table 19: Properties hardware

5.8.2 Implementation ArcGIS Pro

Performance in ArcInfo Workstation is the performance of the current system, the system under test is ArcGIS Desktop, or, eventually, ArcGIS Pro. Little is yet known about how geoprocessing as well as administrative tools are handled in ArcGIS Pro. The performance of the prioritized tool will be measured of the default settings and compared to ArcGIS desktop default settings. As it is only possible to test ArcGIS Pro in the innovation laboratory on the big data computer, ArcGIS Pro should be tested on a fat client later on.

Chapter 6: Results and Conclusions of the Benchmark

6.1 Results Baseline tests

6.1.1 Baseline ArcInfo Workstation

To establish a baseline of current performance of geoprocessing tools in ArcInfo Workstation and to study the effect of the upgrade of the operating system to Windows 7, measurements of geoprocessing tools in ArcInfo Workstation 9 and 10 have been taken. Workstation 9 has been tested on a relatively "old" fat client machine with OS Windows XP and "older" hardware properties to sketch the situation before the upgrade (see section 3.3.2 Fat clients properties). The measurements of Workstation 10 on Windows 7 show the current situation and depict the baseline for performance. Effectively, the old and new fat clients represent at least 3 different performance factors: Hardware resources, Operating System (OS) and GIS software version. ArcInfo Workstation 10 has been tested as well on a Windows XP configuration, to exclude the factor of a higher ArcInfo Workstation version, while the combination hardware and OS still remains. Measurements on Windows XP fat clients show that all geoprocessing tools run slower within the Windows XP environment and Workstation 9, except for an administrative "cursor" tool (in the test situation, an update cursor). The measurements of the combination Windows XP and ArcInfo Workstation 10 show that this combination shows the slowest performance, except for the cursors (short and long version), slower than the combination Windows XP and ArcInfo Workstation 9. The reason behind the slow performance of the cursor on the new fat clients could be a different algorithm execution in Workstation 10, but then the tool should have also performed slower on the old fat client with Workstation 10. Therefore, the reason could also be the operating system, Windows 7.

The graph shows the execution time of the tools in seconds. It is also important to know that the presented tools have been using different input datasets, representative for their use at SN. The union crashed in Windows XP and Windows 7 configurations with the land use file (BBG), therefore this tool has been running with a smaller dataset, the neighbourhood- and district-boundaries (wijk- en buurtgrenzen). Although input datasets are different per tool, the difference between the three configurations shows a clear trend, with one exception only (the cursor). Figure 27 also shows that the performance is also dependent on the combination of software version and version of the operating system, since most combinations of Windows XP and ArcInfo Workstation 10 show a slower execution time than Windows XP and ArcInfo Workstation 9.

During the baseline test in ArcInfo Workstation, the DISSOLVE (on a dataset of 233.153 records) showed the biggest difference in execution time between the different configurations, followed by the UNION (on a small dataset). This result could lead to the assumption that the higher Workstation version is not the cause of the improved performance, but rather the hardware and specifications that influence the DISSOLVE tool more than the other tools, at least with this dataset and DISSOLVE settings.

The values of the execution time is presented in seconds on a logarithmic scale because of the difference between the execution time of the NEAR and the execution time of the remaining tools.

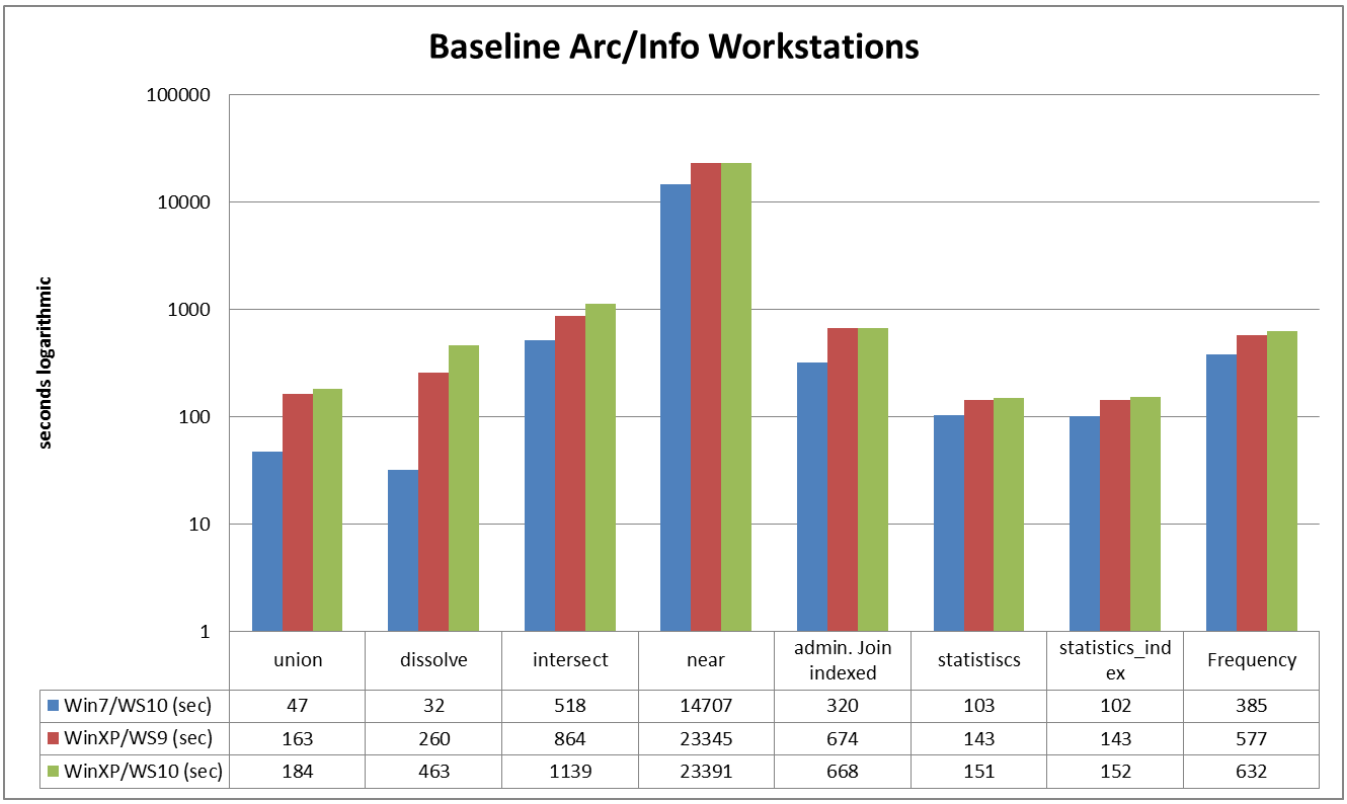


Figure 27: Result (execution time) ArcInfo Workstation baseline test on a selection of geoprocessing and administrative processing tools

6.1.2 Baseline ArcGIS 10.1

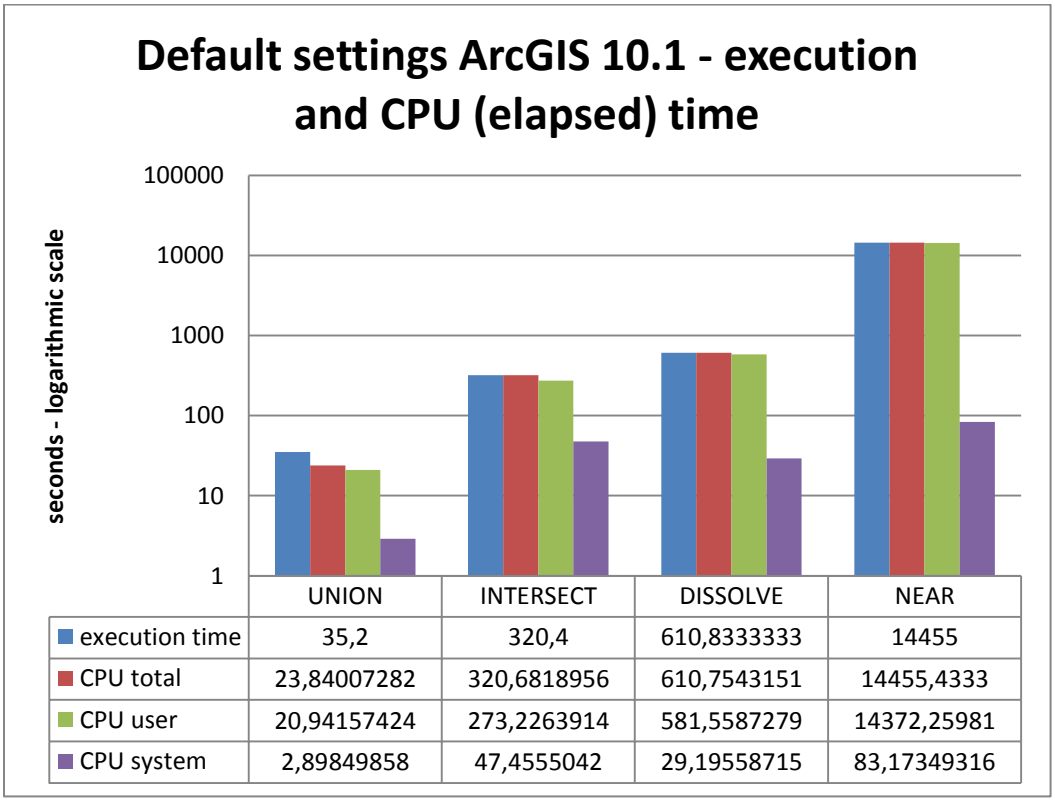


Figure 28: UNION, INTERSECT, DISSOLVE and NEAR execution and CPU time with ArcGIS 10.1 on fat clients Windows 7, default settings

Table 20 shows an overview of the execution times of the same scenario's in Workstation versus Desktop (10.1) in default state. It shows that, using these scenarios, only the dissolve is faster (20 x faster). The other scenario's result in slightly faster execution times.

Tool	Datasets	No. records	Win7/WS10 (sec)	ArcGIS 10.1 (sec)
Dissolve	BBG2008	233.153	**32**	**611**
Intersect	• (enriched road map) nwb_split • (land use file/wood) bos_bbg08	2.826.593 (NWB), 25.157 (BBG_bos)	**518**	**321**
Near	• NWB_split	(NWB), 126.331 (intersect_p)	**14.707**	**14.455**
Union	• NEDWB_08 • NEDWB12	125.70 (WB2008), Xxx	**47**	**35,2**

Table 20: Execution time ArcInfo Workstation versus ArcGIS 10.1 default real workload

Table 21 will show a number of scenario's for administrative processing. These will not be optimized further, but to show the difference between ArcInfo Workstation and ArcGIS 10.1 Desktop (in default settings), these results will provide a more complete overview. The administrative tools are all tested with a large dataset of approximately 8.500.000 records. Except for the JOIN, the other tools (SUMMARY STATISTICS and FREQUENCY) show improvement in execution time in ArcGIS 10.1, without additional optimization factors.

Tool	Datasets	No. records	Win7/WS10 (sec)	ArcGIS 10.1 (sec)
ADMIN. JOIN INDEXED	• House numbers (num2012ba) • vsl2012ba.pull	• (num2012ba) 8.668.742 • (vsl2012ba.pull) 8443296	**320**	**8229**
SUMMARY STATISTICS	House numbers without values "NULL"	8.443.273	**102**	**52**
FREQUENCY	House numbers	8.668.742	**385**	**270**

Table 21: Administrative processing ArcInfo Workstation vs ArcGIS Desktop 10.1 default settings

6.2 Results Scalability and data size

6.2.1 Results synthetic data INTERSECT AND UNION

As explained in chapter 5.4, a very large input dataset of 150.100.000 fishnet grid polygons will be used for the INTERSECT and the UNION in an overlay with a second dataset that increases in in size (number of records): 100.000, 500.000, 1.000.000, 5.000.000, 10.000.000 and 50.000. The last two runs of 10.000.000 and 50.000.000 have not been tested for the UNION because of the long execution time in combination of lack in time resources. The execution time results of the INTERSECT does not show a linear line. The run of 1.000.000 versus 150.100.000 records shows a lower execution time than the scenario of 100.000 versus 150.100.000 records, whereas the run using 50.000.000 versus 150.100.000 records show a higher value than expected within a linear line. These scenarios have been repeated once more and resulted in the same values.

Figure 29 shows a big difference between the execution time of the UNION and the INTERSECT, although these tools in fact follow the same steps during the overlay process, except for writing the new output: the INTERSECT output contains only the overlapping geometry, UNION output contains the overlapping plus remaining geometry, so more has to be written to disk. In this case, the combination with such a large input dataset (150.000.100 records) combined with a seconds input dataset of from 100.000 records, for example, results in I/O time needed for the UNION has to read 150.000.100 and to write app. 150.200.000 features, whereas the INTERSECT needs I/O time to read the same number of input feature but write only app. 200.000 features. Moreover, the second input dataset is very small compared to the first input dataset and the variation in size is very small. Therefore, the execution time hardly changes and the system time percentage stays at app. 40% (see Figure 30).

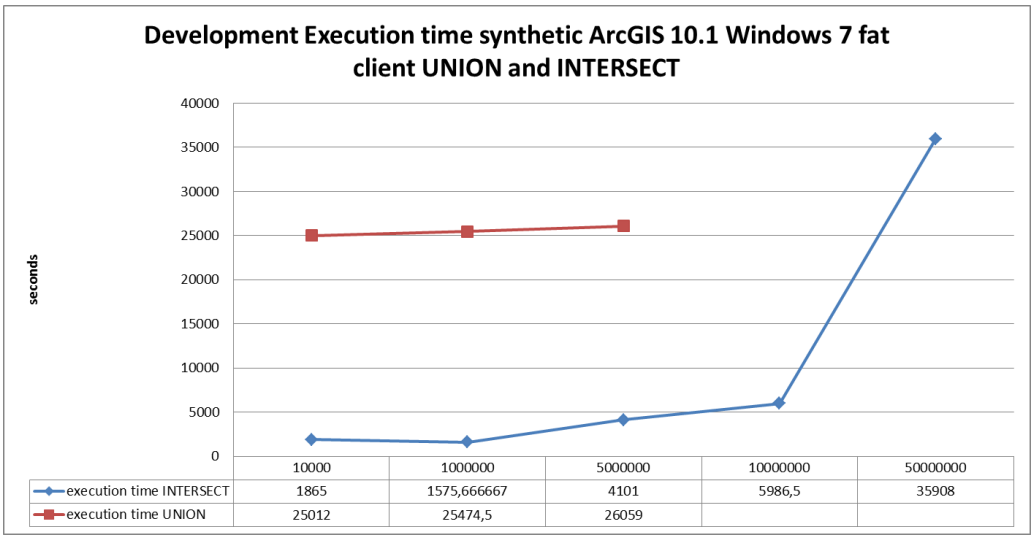


Figure 29: INTERSECT synthetic data using a constant input dataset of 150.000.100 records and a second input dataset with sizes from 100.000 to 50.000.000 records

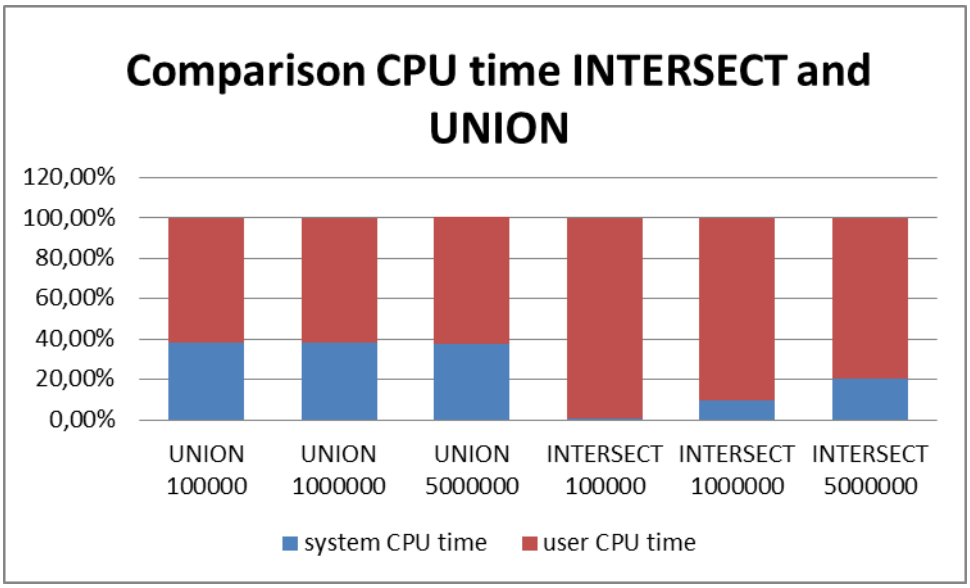


Figure 30: CPU system and user time ratio INTERSECT and UNION - ArcGIS 10.1 Windows 7 fat client

6.2.2 Results synthetic data DISSOLVE

The results shown in Figure 31 and Figure 32 show a linear development of the execution time related to the number of records, up to the maximum number of 50.000.000 records. A higher number has not been included in the test because the creation of the feature class already takes a long time. The comparison with ArcInfo Workstation is only partially possible: The DISSOLVE has been tested with coverages of 100.000, 500.000 and 1.000.000 records because of the size limitations of the coverage. For the mentioned sizes, the DISSOLVE has a shorter execution time in comparison with the file geodatabase (Figure 33: Comparison DISSOLVE synthetic data in ArcInfo Workstation and ArcGIS Desktop 10.1). The development of the results is linear in both cases.

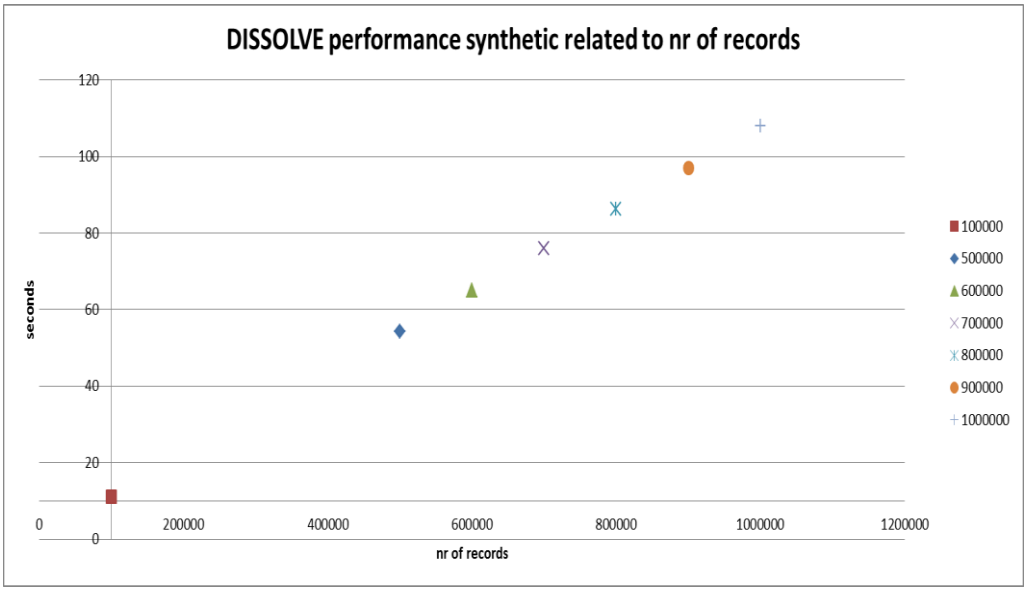


Figure 31: Results feature class up to 10.000.000 records

The results of the DISSOLVE have been visualized in two figures to be able to show the linear development with datasets up to 1.000.000 and the development with datasets up to 50.000.000 records.

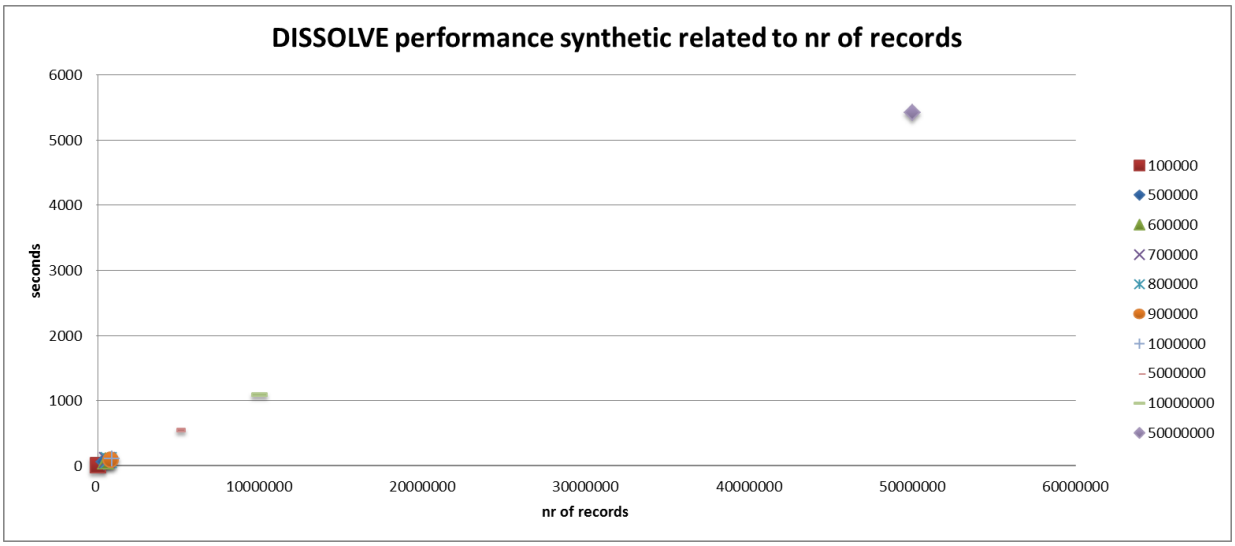


Figure 32: Results feature class up to 50.000.000 records

Considering Figure 33, it is also interesting to note that the coverages are showing better performance for 100.000, 500.000 and 1.000.000 records, but the acceleration of the coverage is “only” between 4,4 and 3,5,

compared to 20 for the real life data (see Table 20: Execution time ArcInfo Workstation versus ArcGIS 10.1 default real workload). The stored topology with the principle of contiguity provides more advantage for the coverage for real data than the synthetic data, since the real data contain more vertices. Still, the creation of input coverages shows limits of the scalability of ArcInfo Workstation.

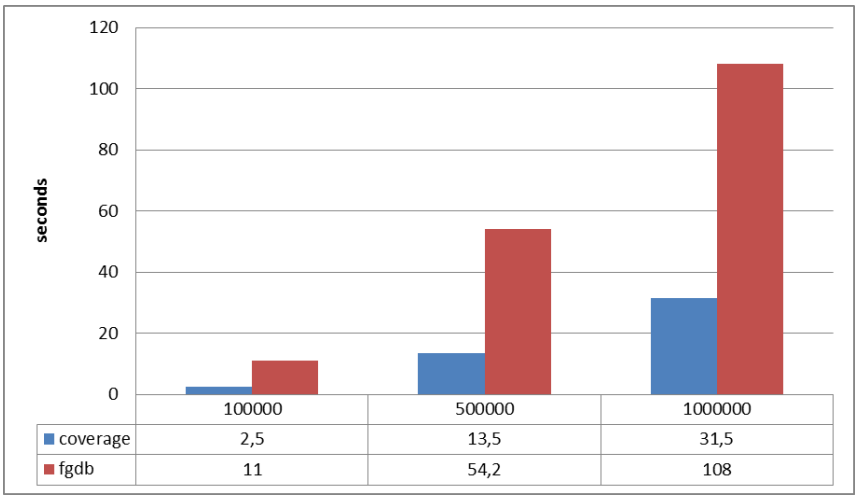


Figure 33: Comparison DISSOLVE synthetic data in ArcInfo Workstation and ArcGIS Desktop 10.1 for 100.000, 500.000 and 1.000.000 records

6.2.3 Results synthetic data NEAR

The NEAR shows a linear curve as shown in Figure 34 after plotting the execution time results. Due to the long execution time of the 500.000 records, further upscaling has not been attempted.

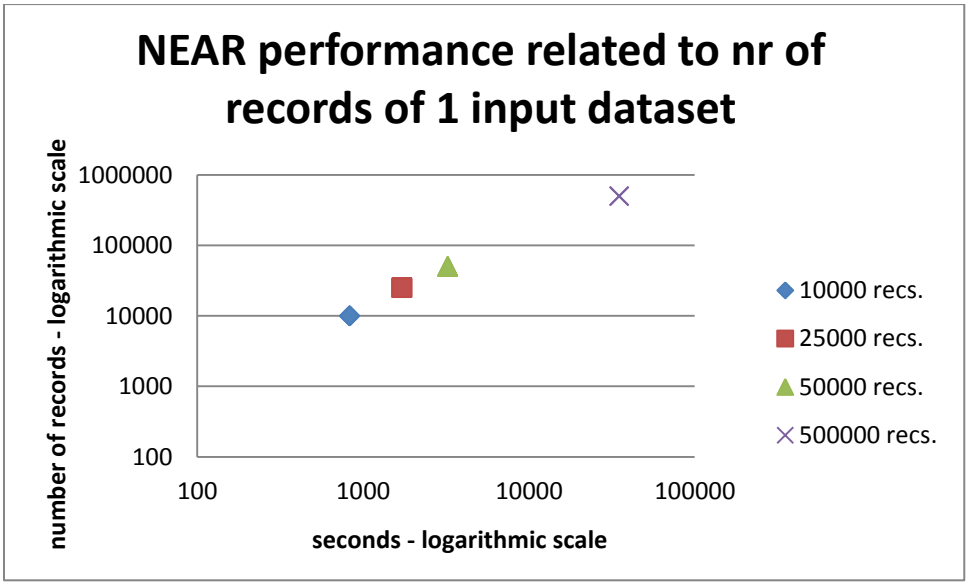


Figure 34: NEAR synthetic - linear curve

The following chart clearly show that most of the CPU time is used by the actual calculation: the system time only shows a slight rise compared to the linear rise of the CPU user time.

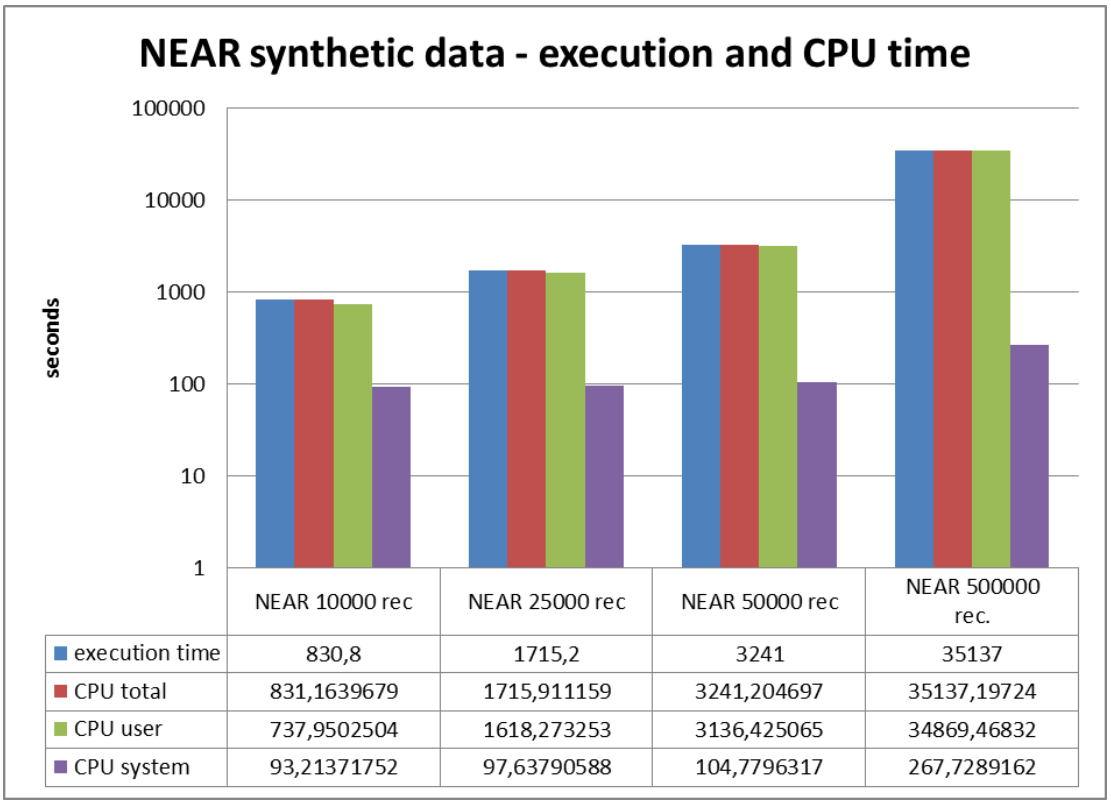
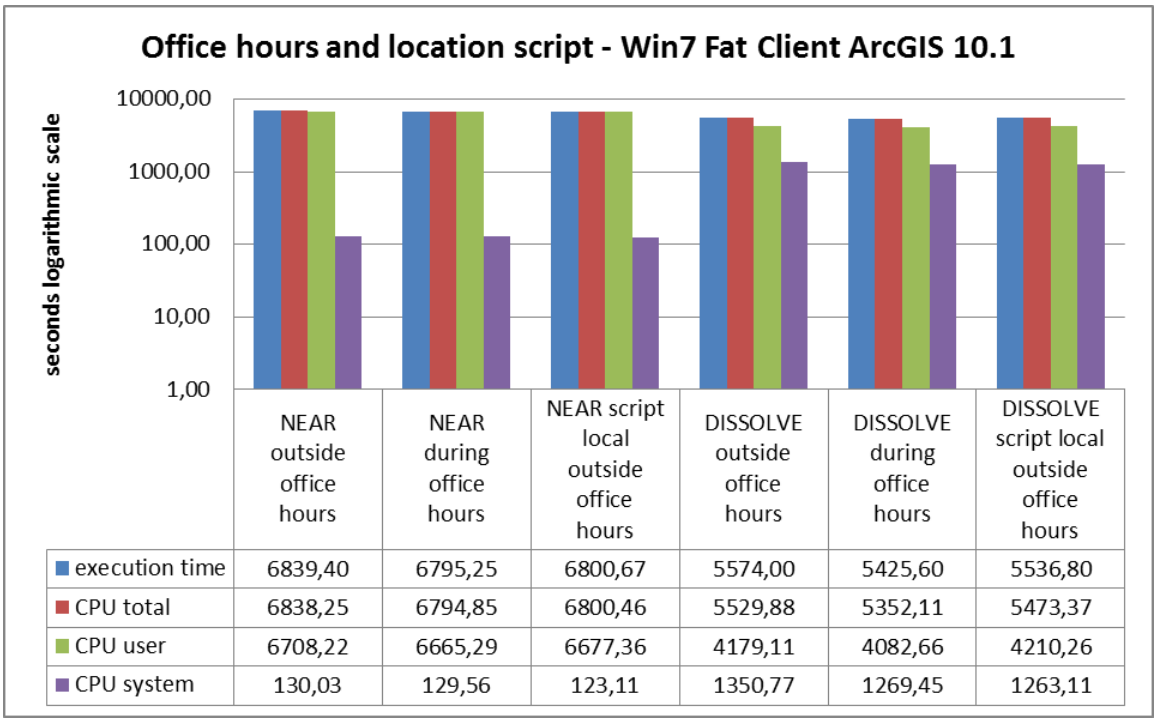


Figure 35: NEAR synthetic data - execution and CPU time

6.3 Exclusion of noise due to network interference or location of input and output data in workspace

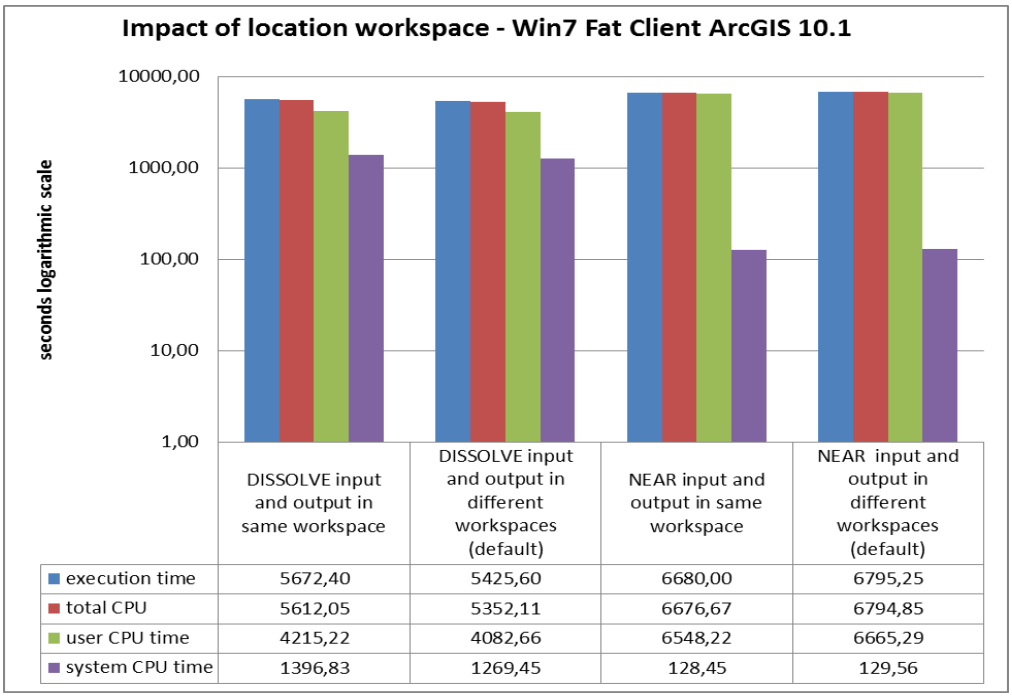
The results of the NEAR show practically no difference in execution time and distribution of CPU user and system time. For this scenario, no effect of network interference has been measured. The same can be said for the DISSOLVE. Similar to the NEAR, the execution and CPU time is slightly longer than during office hours. There is no impact of less or more network traffic over the network. One of the reasons could be the fact that only one tool is run at the same time, instead of a sequence of tools.



	NEAR outside office hours	NEAR during office hours	NEAR script local outside office hours	DISSOLVE outside office hours	DISSOLVE during office hours	DISSOLVE script local outside office hours
execution time	6839,40	6795,25	6800,67	5574,00	5425,60	5536,80
CPU total	6838,25	6794,85	6800,46	5529,88	5352,11	5473,37
CPU user	6708,22	6665,29	6677,36	4179,11	4082,66	4210,26
CPU system	130,03	129,56	123,11	1350,77	1269,45	1263,11

Figure 36: Execution and CPU time NEAR outside and during office hours and script local outside office hours (ArcGIS 101 Fat Client)

For the location of input and output data, no significant difference could be measured, as expected. Storage of the input data and output data in the same file geodatabase led to similar values as storage in different file geodatabases:



	DISSOLVE input and output in same workspace	DISSOLVE input and output in different workspaces (default)	NEAR input and output in same workspace	NEAR input and output in different workspaces (default)
execution time	5672,40	5425,60	6680,00	6795,25
total CPU	5612,05	5352,11	6676,67	6794,85
user CPU time	4215,22	4082,66	6548,22	6665,29
system CPU time	1396,83	1269,45	128,45	129,56

Figure 37: Impact of same or different workspace (fgdb) for in- and output data

The results show almost no difference in performance improvement for DISSOLVE and NEAR: The difference either way between keeping the input and output data in the same or in different workspaces is very small,

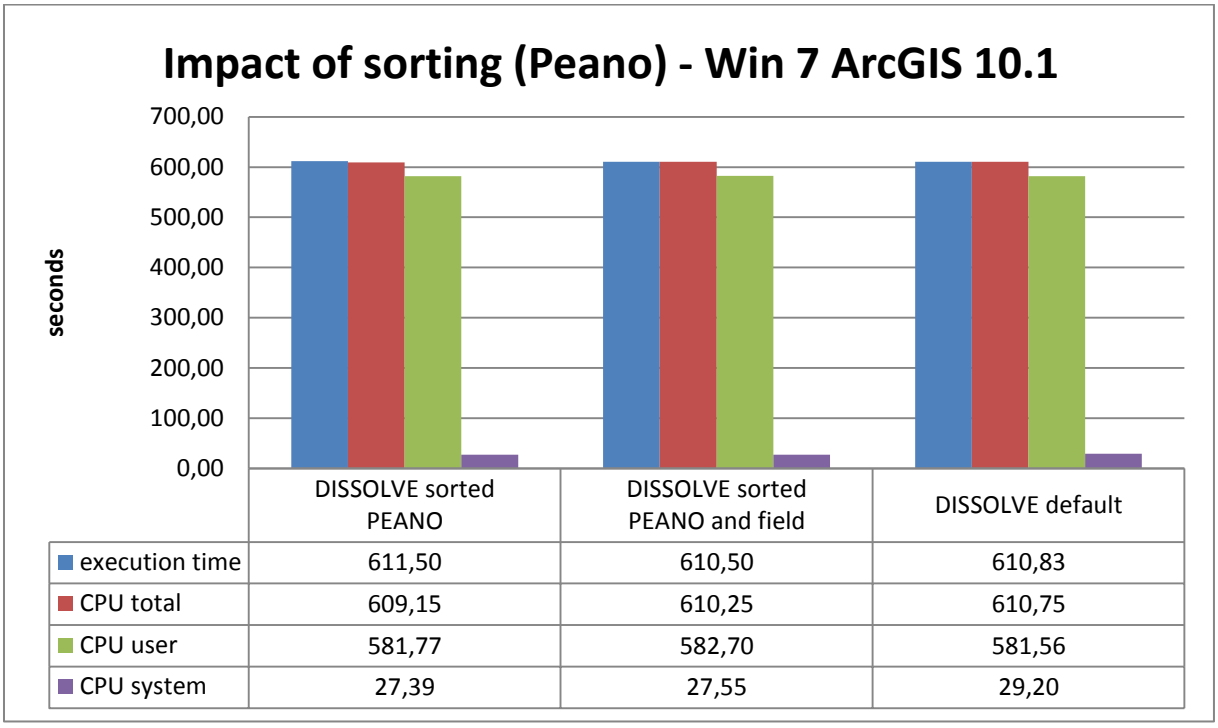
related to the execution time, and therefore not useful to implement as an optimization factor. No significant difference could be detected for network interference in terms of script location and running a tool during a time period that is expected to have less network activity. This is not a surprising result given the fact that the data are stored locally, as well as the I/O process and the calculation. There is no difference between the ratio of CPU user time and system time, compared to the default settings. The same is true for the workspace factor: The acceleration is only 0,02 for the NEAR and a slightly slower execution time for the DISSOLVE (0,05). The NEAR has been assumed to have slightly better results, because the NEAR distances are added to one the input datasets. These results lead to the conclusion that "noise" via the network and workspace can be excluded.

6.4 Sorting and indexing

Goal of this benchmark component is to test the impact of the available means in ArcGIS Desktop 10.1 to organize the data in an efficient way. As explained in CHAPTER 2, the possibilities to apply spatial access methods in ArcGIS desktop are limited.

6.4.1 Sorting (Peano)

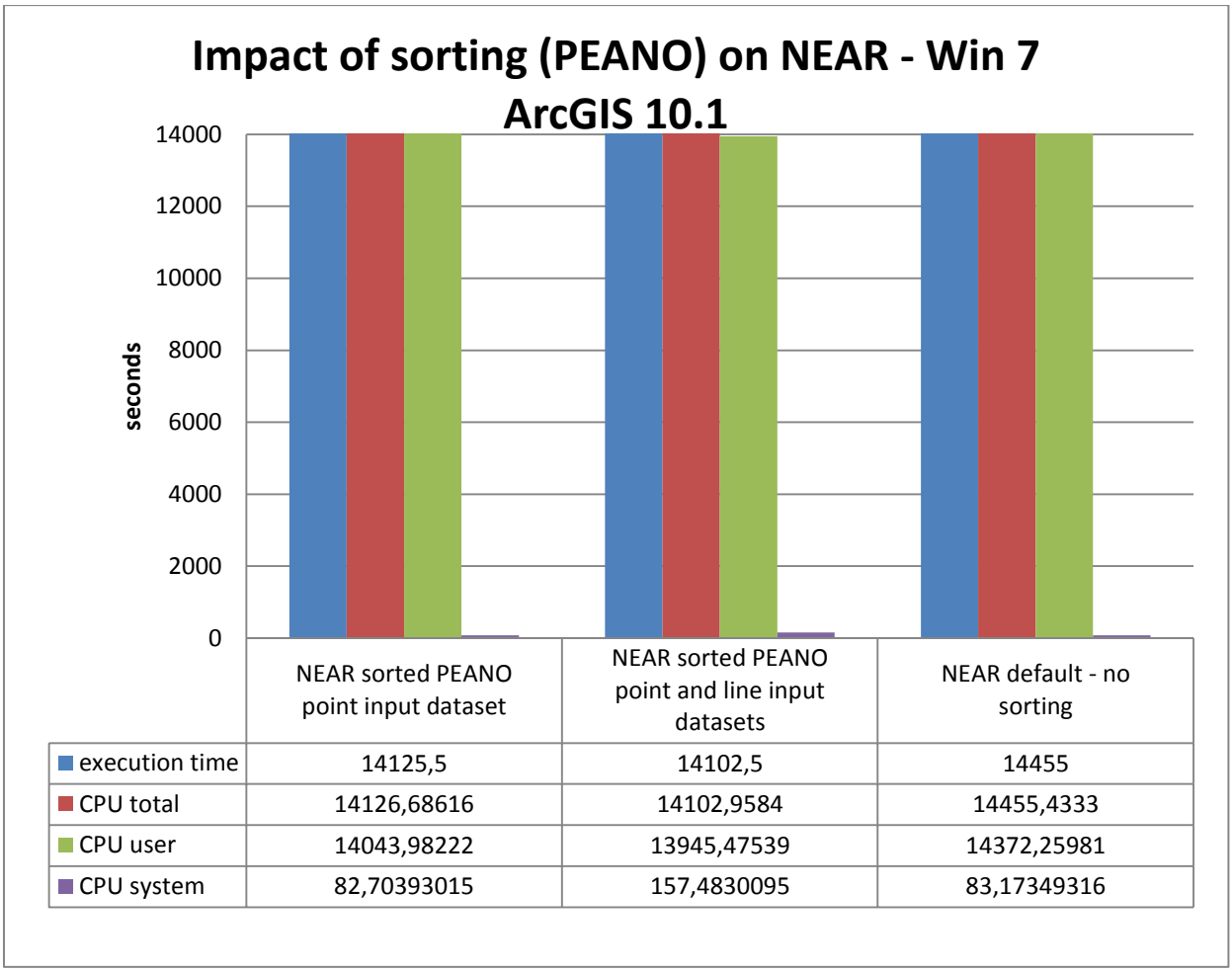
The application of sorting shows no improvement in performance for the tool DISSOLVE. The ratio between CPU user and system time remains the same. The combination of Peano sorting and field (row prime) sorting shows a minimal difference with the default solution.



	DISSOLVE sorted PEANO	DISSOLVE sorted PEANO and field	DISSOLVE default
execution time	611,50	610,50	610,83
CPU total	609,15	610,25	610,75
CPU user	581,77	582,70	581,56
CPU system	27,39	27,55	29,20

Figure 38: Impact of sorting (Peano with and without combination of field sorting) on DISSOLVE (real data)

The same can be said for the NEAR: The sorting of both inputs seems to provide the best result, although the improvement is very small and looking at the original values that are the basis for the medium value, the medium of the sorting operation involving only the point input has been derived from values 14198 and 14053 seconds, whereas the sorting of both input datasets resulted in values of 14109 and 14096 seconds.



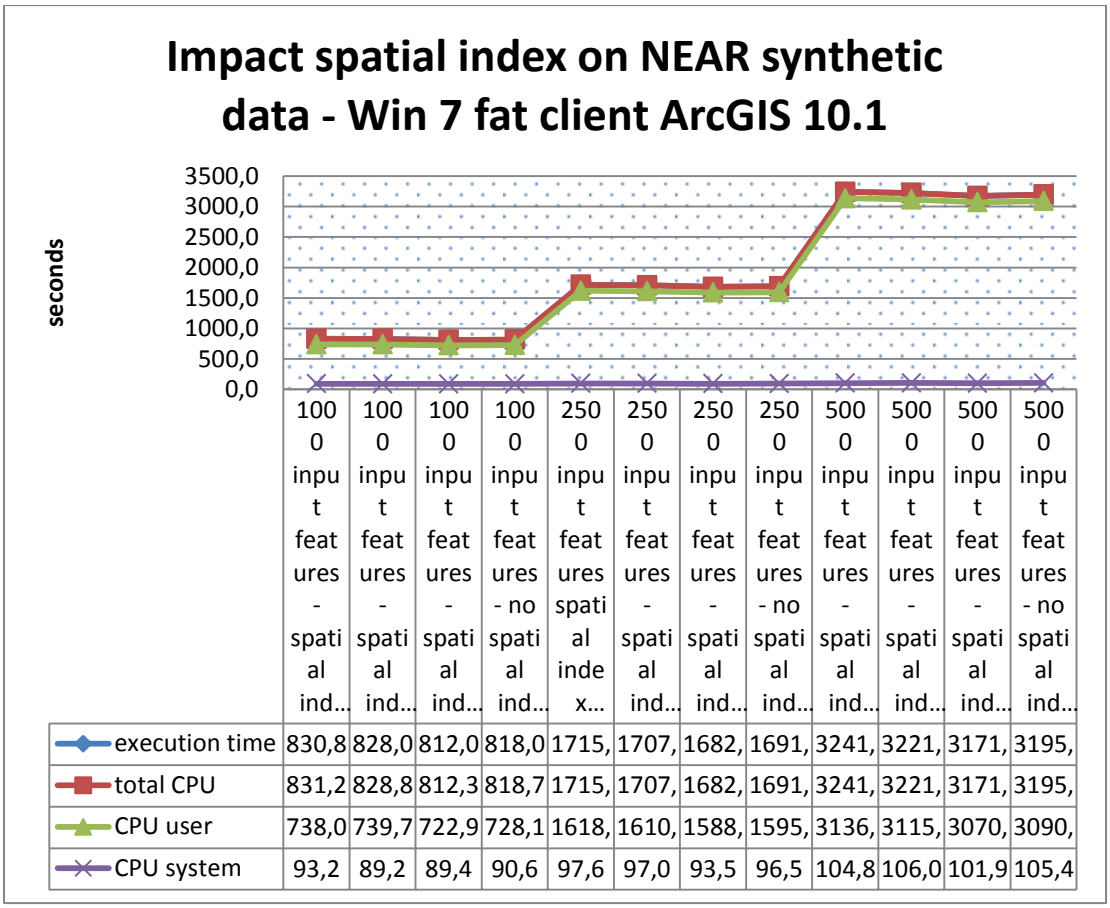
	NEAR sorted PEANO point input dataset	NEAR sorted PEANO point and line input datasets	NEAR default - no sorting
execution time	14125,5	14102,5	14455
CPU total	14126,68616	14102,9584	14455,4333
CPU user	14043,98222	13945,47539	14372,25981
CPU system	82,70393015	157,4830095	83,17349316

Figure 39: Impact of sorting (Peano) on NEAR with options of sorting one or both input datasets

The space-filling Peano curve subdivides the space into sub tiles, therefore sorting both input datasets are probably subdivided into the same sub tiles, and consequently searching for the nearest neighbour will be within the "common" tile first.

6.4.2 Spatial Index

Figure **40** shows that no significant impact can be detected for the NEAR calculation from 10000, 2500 or 5000 input points to the polyline dataset. Even the option of "no spatial index" does not make a difference. This can be stated for the execution time as well as for the CPU user time. The CPU system timeline only shows a very slight slope. The difference between the number of input features seems to be too small for higher I/O activity, but the computation of NEAR features takes more time. This could mean that more calculation capacity from the CPU is needed, just because of the higher number of records, whereas the application of the spatial index does not reduce the needed calculation capacity. The spatial index does not lead to higher I/O activity, as the difference between the values of CPU system time of the runs with or without spatial index is very small.



	1000 input features - spatial ind...	1000 input features - spatial ind...	1000 input features - spatial ind...	1000 input features - no spatial ind...	2500 input features spatial index...	2500 input features - spatial ind...	2500 input features - spatial ind...	2500 input features - no spatial ind...	5000 input features - spatial ind...	5000 input features - spatial ind...	5000 input features - spatial ind...	5000 input features - no spatial ind...
execution time	830,8	828,0	812,0	818,0	1715,	1707,	1682,	1691,	3241,	3221,	3171,	3195,
total CPU	831,2	828,8	812,3	818,7	1715,	1707,	1682,	1691,	3241,	3221,	3171,	3195,
CPU user	738,0	739,7	722,9	728,1	1618,	1610,	1588,	1595,	3136,	3115,	3070,	3090,
CPU system	93,2	89,2	89,4	90,6	97,6	97,0	93,5	96,5	104,8	106,0	101,9	105,4

Figure 40: Impact of spatial index on NEAR in ArcGIS 10.1

6.4.3 Parallel Processing

The parallel processing environment setting is a parameter that can be adapted per tool. Leaving the parameter empty will let the tool assign the CPU tasks. Setting the factor to 50%, 75% or 100% does lead to a decline in performance for the DISSOLVE and almost no improvement in performance for the NEAR. But also the distribution of the CPU percentage does not match with the factor (Table 22): For example, for 100%, a more even distribution between the cores would be expected, e.g. 4 x 25 %.

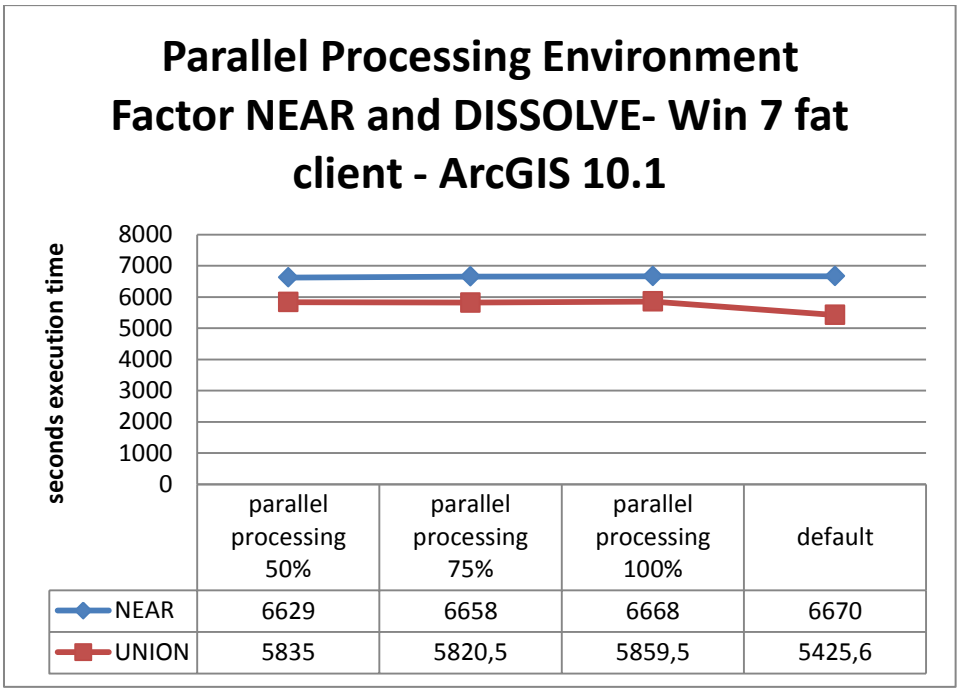


Figure 41: Impact of adapting parallel processing environment setting with DISSOLVE and NEAR

Run	CPU percentage	Core1	Core2	Core3	Core4
1	50%	3,9	5	45,9	32,2
2	50%	8,1	4,7	45,7	40,2
1	75%	8,2	0	47,2	37,8
2	75%	9,3	6,3	38,8	37,9
1	100%	5,7	5,5	46,2	34,9
2	100%	7,3	4,3	37,3	49,9

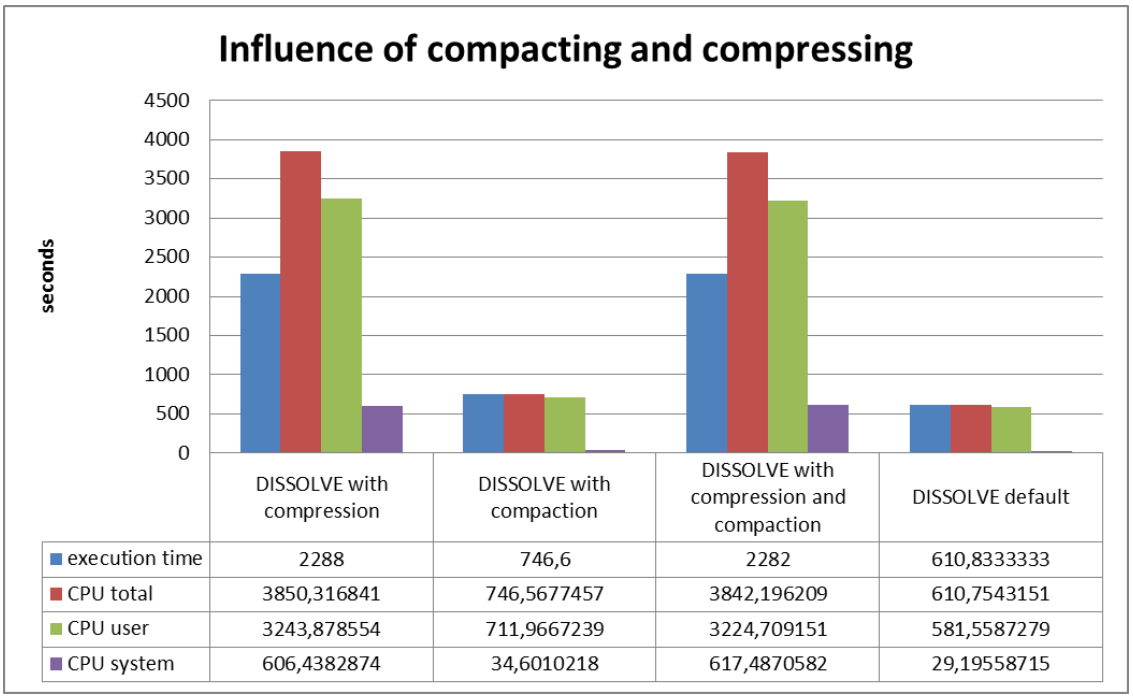
Table 22: CPU percentages - after adaption of parallel processing environment setting

Setting the parallel processing environment does not seem to work well in terms of performance but also in terms of actual distribution of computation power among the cores. The implementation of the parallel processing environment does not provide the desired result for the DISSOLVE. However, it would still take more time to experiment with parallel processing with the use of other tools but also with parallel processing via Python scripting.

6.5 Compression and compaction

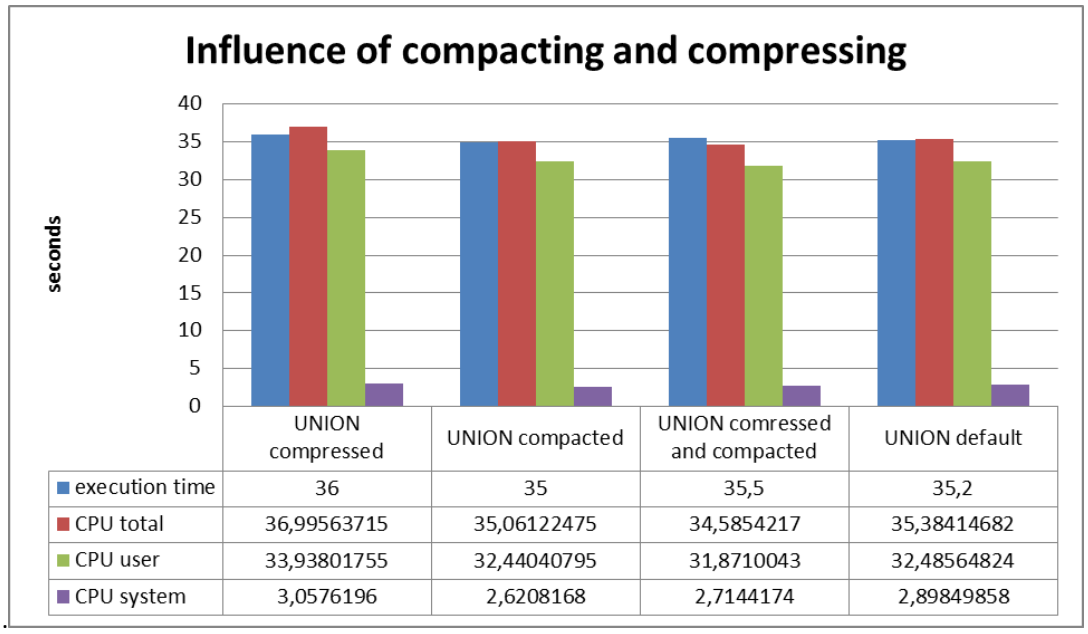
Although expectations of these optimization methods have not been high, the results are still surprising: with the compressed dataset, the DISSOLVE runs 3,75 slower than the default setting, also in combination with a compaction. The results also show a higher CPU time: the total CPU time as well as the CPU user time is higher than the execution time, meaning that more CPU capacity is needed than available for the compressed data. It also shows that the I/O proces is not the bottleneck for the operation but thecomputation. After examination of the CPU percentages, all cores point at a CPU use of 170%, compared to app. 100% with the default or compacted data. The CPU time represents the sum of all cores. A valid reason might be the need to de compress the data first prior to the actual processing, which adds to the computational workload, although the compressed data can be accessed directly according to (ESRI, 2013b). The UNION for compacted and compressed shows a completely different scenario: No difference between all variations. The scenarios with compression are not slower than the other scenario's. Additionally, no difference can be detected in CPU user

and system time distribution. Possible reasons are: The datasets are small and have not been edited frequently; the extra workload of uncompressing the data is scarce, although some difference should have been noticed.



	DISSOLVE with compression	DISSOLVE with compaction	DISSOLVE with compression and compaction	DISSOLVE default
execution time	2288	746,6	2282	610,8333333
CPU total	3850,316841	746,5677457	3842,196209	610,7543151
CPU user	3243,878554	711,9667239	3224,709151	581,5587279
CPU system	606,4382874	34,6010218	617,4870582	29,19558715

Figure 42: Impact of compacting and compressing on DISSOLVE (Win 7 fat client)



	UNION compressed	UNION compacted	UNION comressed and compacted	UNION default
execution time	36	35	35,5	35,2
CPU total	36,99563715	35,06122475	34,5854217	35,38414682
CPU user	33,93801755	32,44040795	31,8710043	32,48564824
CPU system	3,0576196	2,6208168	2,7144174	2,89849858

Figure 43: Influence of compacting and compressing on the UNION (Win 7 fat client)

Compacting shows no influence on the NEAR operation, as expected. The input datasets have not been updated frequently at Spatial Statistics, therefore compacting does not lead to a dataset that is smaller in size and more efficiently stored. Compressing could not be tested for the NEAR because the compression of one of the input datasets (the NWB), resulted in an error.

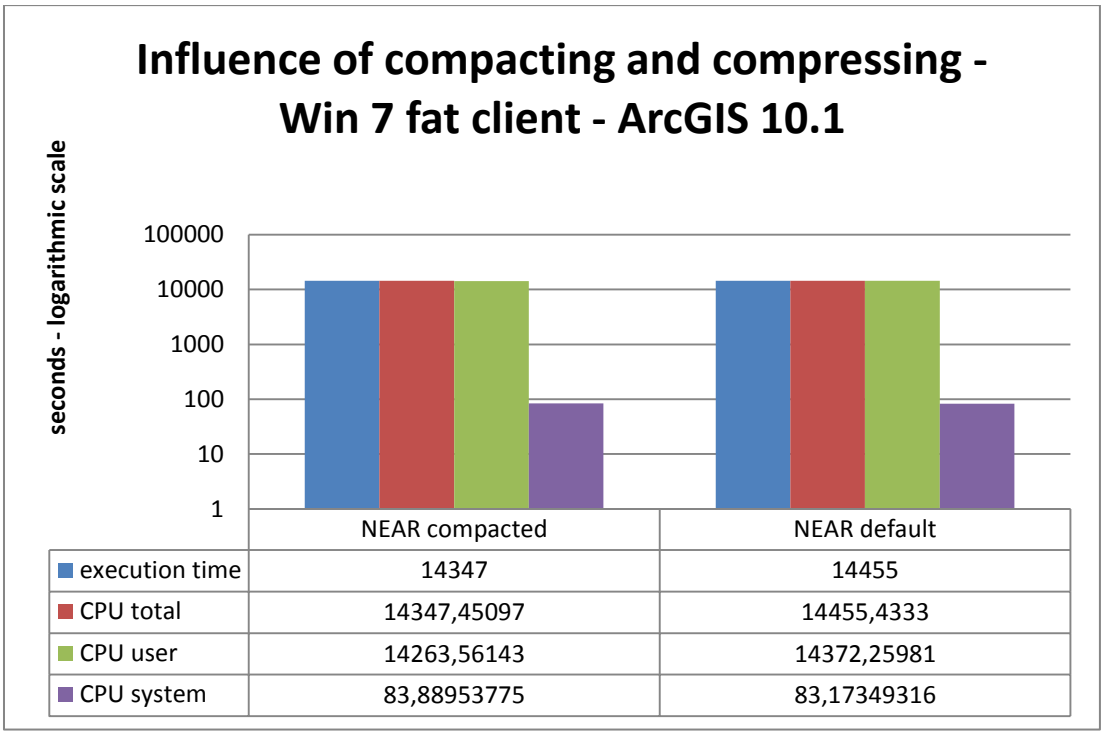
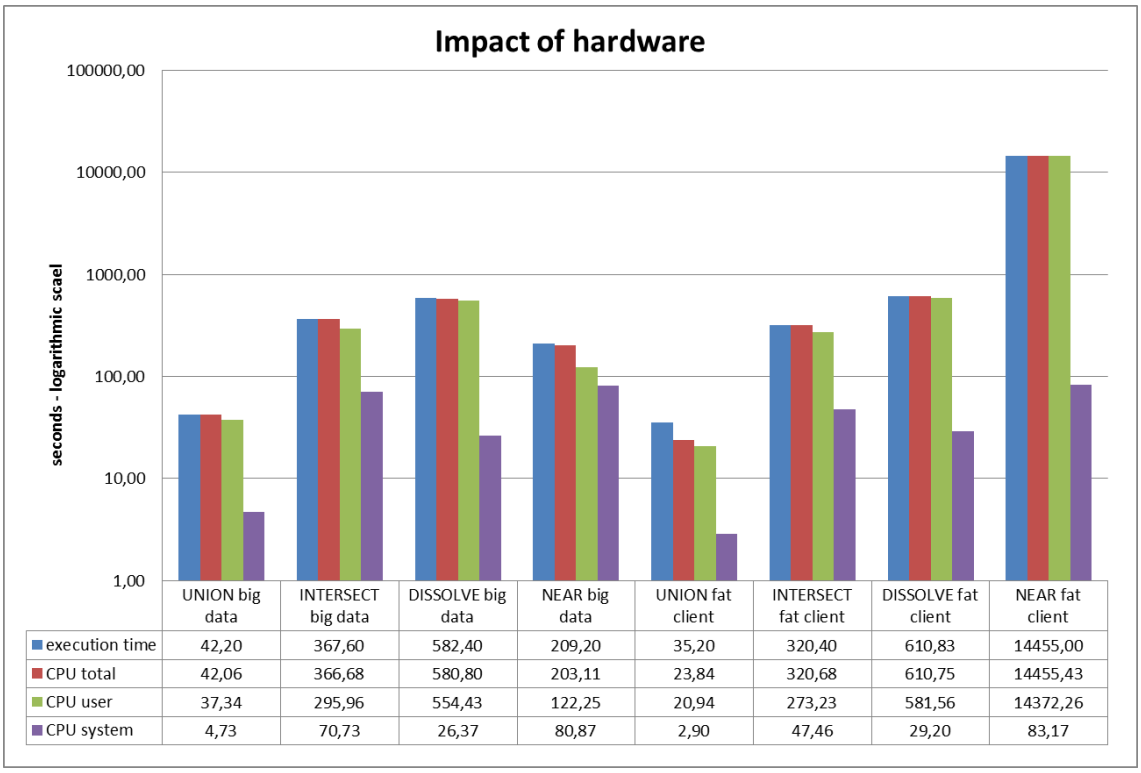


Figure 44: Impact of compacting on NEAR

6.6 Hardware configuration and ArcGIS Pro

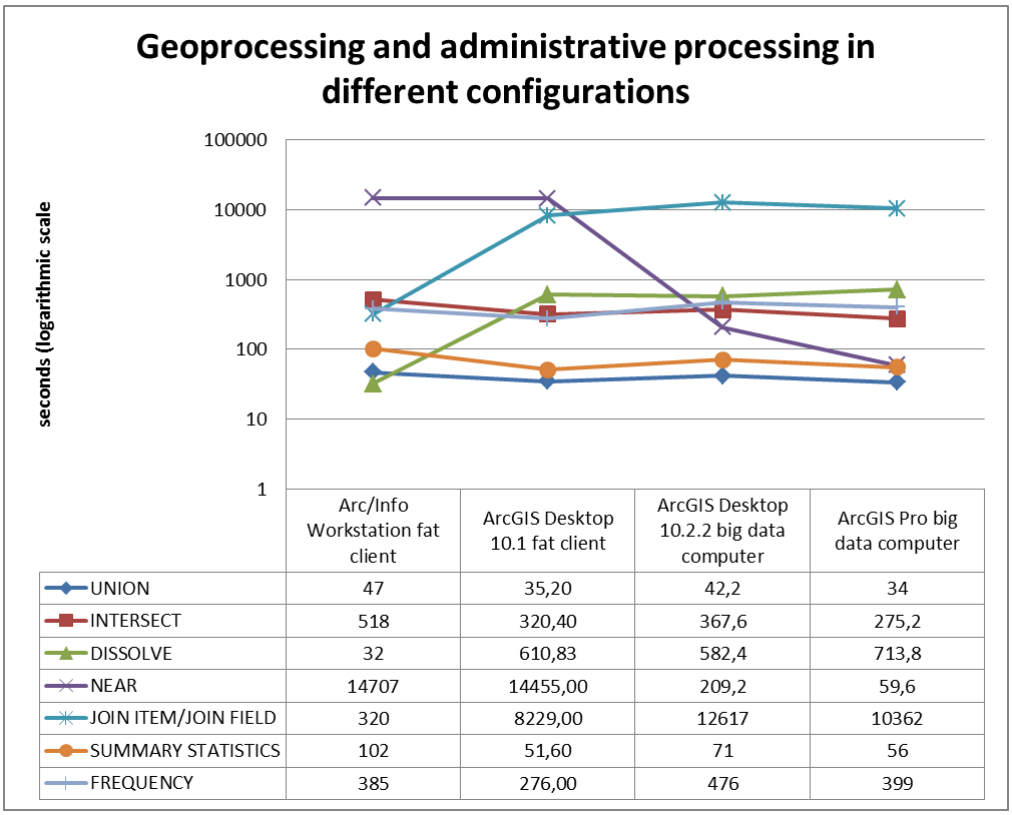
It is difficult to compare the benchmark results of the workload in ArcGIS 10.1 on the normal fat client to the results with a higher version on a computer with better resources. Still, the results of the benchmark tests combining the big data configuration with ArcGIS 10.2.2 (instead of ArcGIS 10.1) show remarkable developments: No speed up, even slightly slower results for the UNION and the INTERSECT, only slight speed up of the DISSOLVE, however a speed up factor of 69 and app. 71 for execution time and CPU time (total) for the NEAR. The values for CPU user and system time for the NEAR also show (Figure 45) that the value for CPU system time stays at the same level in both configurations whereas the CPU user time shows a sharp decline. Correlated with the execution time and the redevelopment of the NEAR, it points at a smaller calculation capacity that is needed, due to the improved, more efficient algorithm of the GENERATE NEAR TABLE and NEAR tools (ESRI, 2014d), which has also been referred to in section 2.1 ESRI software development).The other geoprocessing tools did not show any improvement, rather a slight decline for the INTERSECT and the UNION and a very slight improvement for DISSOLVE. The big data resources as well as the new software version did not lead towards improved performance for these tools.



	UNION big data	INTERSECT big data	DISSOLVE big data	NEAR big data	UNION fat client	INTERSECT fat client	DISSOLVE fat client	NEAR fat client
execution time	42,20	367,60	582,40	209,20	35,20	320,40	610,83	14455,00
CPU total	42,06	366,68	580,80	203,11	23,84	320,68	610,75	14455,43
CPU user	37,34	295,96	554,43	122,25	20,94	273,23	581,56	14372,26
CPU system	4,73	70,73	26,37	80,87	2,90	47,46	29,20	83,17

Figure 45: Comparison of fat client (ArcGIS 10.1/Windows7) and big data computer (ArcGIS 10.2/Windows 7) performance

The next step of the benchmark were the performance tests with the real data workload on ArcGIS Pro on the big data computer. Instead of showing separate results for ArcGIS Pro, the execution time results of all real data workload scenarios (without other factors) have been visualized in Figure 46. It shows different outcomes per tool: UNION (using two smaller input datasets) and INTERSECT show improvement from ArcInfo Workstation to ArcGIS 10.1 on the fat client, a slightly slower execution time on the big data computer with ArcGIS Desktop 10.2 and a faster time with ArcGIS Pro on the big data computer. DISSOLVE clearly shows a considerable slowdown of execution time in all higher desktop configurations, surprisingly with ArcGIS Pro on the big data computer as the slowest configuration. The NEAR shows a big improvement from ArcGIS 10.1 on the fat clients to ArcGIS 10.2.2 on the big data computer and from ArcGIS 10.2.2 to ArcGIS Pro. The administrative processing tools show no improvement of the result in ArcInfo Workstation, except for the SUMMARY STATISTICS. Especially the JOIN FIELD shows a dramatic decline in performance from ArcInfo Workstation, first from Workstation to Desktop on the fat client, but also during the consequent steps on the big data computer. The execution time has been the slowest in ArcGIS 10.2.2 on the big data computer. Given the performance requirement of Spatial Statistics that the execution time of the tools in Desktop should not be slower than in ArcInfo Workstation, we can only state that this requirement cannot be met in all cases. Some tools perform better in ArcGIS Desktop 10.1 already, but don't show further improvement with higher hardware resources (UNION, INTERSECT, SUMMARY STATISTICS and FREQUENCY perform even slower in ArcGIS 10.2.2). ArcGIS Pro shows improvement compared to ArcGIS 10.2.2 for the same tools, but not faster execution time than ArcGIS 10.1, except for the NEAR and to a lesser degree the INTERSECT and the UNION (for the UNION it could be neglectible, given the small difference).



	Arc/Info Workstation fat client	ArcGIS Desktop 10.1 fat client	ArcGIS Desktop 10.2.2 big data computer	ArcGIS Pro big data computer
UNION	47	35,20	42,2	34
INTERSECT	518	320,40	367,6	275,2
DISSOLVE	32	610,83	582,4	713,8
NEAR	14707	14455,00	209,2	59,6
JOIN ITEM/JOIN FIELD	320	8229,00	12617	10362
SUMMARY STATISTICS	102	51,60	71	56
FREQUENCY	385	276,00	476	399

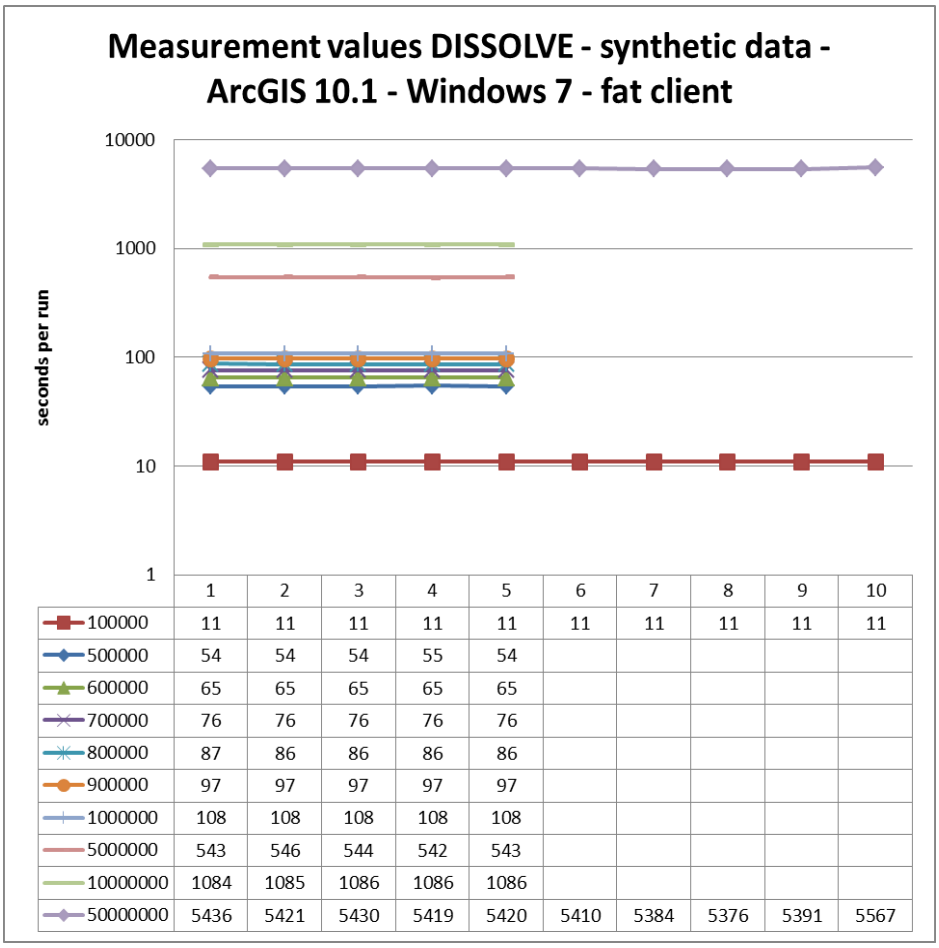
Figure 46: Execution time of geoprocessing tools in the configurations ArcInfo Workstation/fat client, ArcGIS 10.1/fat client, ArcGIS 10.2/big data and ArcGIS Pro/big data. Real datasets re used of different sizes.

6.7 Validation

6.7.1 Analysis of the performance measurements

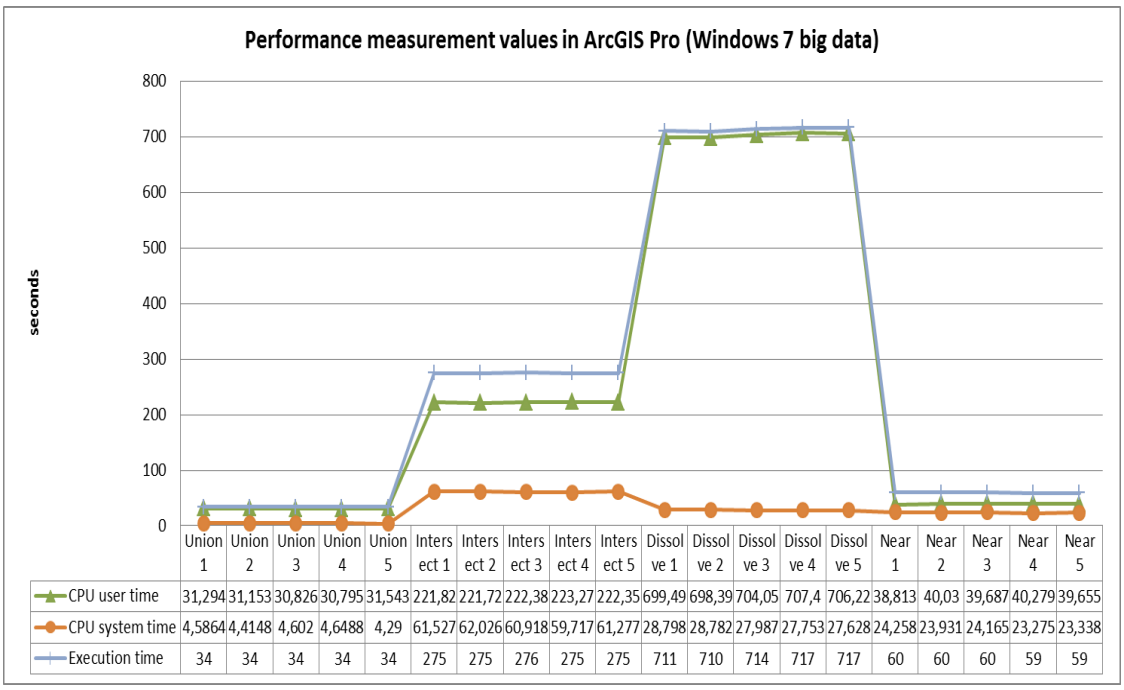
The measurement results of the execution time in seconds have been very stable: at first, 5 runs per factor/tool combination have been executed, after discovering the stability of the measurements; the runs have been reduced to 2, especially for the scenarios with a long execution time. The measurements have shown stable results in almost all combinations. It is also remarkable that the first runs are not necessarily slower than the later runs. This is visible in the table accompanying the chart of Figure 47. The measurements of for 100.000 and 50.000.000 have been recorded 10 times.

Figure 48 also shows continuity of the execution time and CPU values for geoprocessing tools run with ArcGIS Pro on the (Windows 7) big data computer. The chart also shows clearly low CPU system time values for the DISSOLVE.



	1	2	3	4	5	6	7	8	9	10
100000	11	11	11	11	11	11	11	11	11	11
500000	54	54	54	55	54					
600000	65	65	65	65	65					
700000	76	76	76	76	76					
800000	87	86	86	86	86					
900000	97	97	97	97	97					
1000000	108	108	108	108	108					
5000000	543	546	544	542	543					
10000000	1084	1085	1086	1086	1086					
50000000	5436	5421	5430	5419	5420	5410	5384	5376	5391	5567

Figure 47: Measurement values execution time DISSOLVE - synthetic data - ArcGIS 10.1 - Windows 7- fat client



	Union 1	Union 2	Union 3	Union 4	Union 5	Intersect 1	Intersect 2	Intersect 3	Intersect 4	Intersect 5	Dissolve 1	Dissolve 2	Dissolve 3	Dissolve 4	Dissolve 5	Near 1	Near 2	Near 3	Near 4	Near 5
CPU user time	31,294	31,153	30,826	30,795	31,543	221,82	221,72	222,38	223,27	222,35	699,49	698,39	704,05	707,4	706,22	38,813	40,03	39,687	40,279	39,655
CPU system time	4,5864	4,4148	4,602	4,6488	4,29	61,527	62,026	60,918	59,717	61,277	28,798	28,782	27,987	27,753	27,628	24,258	23,931	24,165	23,275	23,338
Execution time	34	34	34	34	34	275	275	276	275	275	711	710	714	717	717	60	60	60	59	59

Figure 48: Performance measurement values in ArcGIS Pro - Windows 7 big data computer

6.7.2 Analysis of the output data

Approximately 350 scenario runs have been executed in total on 4 configurations (including all four configurations generating almost the same number of output files. Because of time constraints, these output files could not be checked on topology, or consistency of field values. Therefore, only the number of records has been checked with each run. For most of the times, the number of records of the output file has resulted in the same number. Only in 2 cases, irregularities were noticed: with compressed files and, the INTERSECT and the DISSOLVE. An error report has been sent for both cases (see Appendix 6: Error Report ESRI)

Compression in ArcGIS 10.1 on the fat client, Windows 7

After checking the number of records for the UNION in the output datasets, it occurred that the output datasets that have been compressed returned 37.278 records, whereas the default setting and the compacted datasets yielded 37.280 records.

The DISSOLVE scenario returned the following results for the compression only: The 4th run suddenly shows 13 records more. This is very curious because it is the same scenario: DISSOLVE with a compressed input file with the configuration ArcGIS 10.1 on the fat client of Spatial Statistics (Operating System Windows 7). It does not result in a higher execution time for the 4th run.

result2_0_3_2_3_0_1_1_0_1_0_1_1_loopnr_1	150074
result2_0_3_2_3_0_1_1_0_1_0_1_1_loopnr_2	150074
result2_0_3_2_3_0_1_1_0_1_0_1_1_loopnr_3	150074
result2_0_3_2_3_0_1_1_0_1_0_1_1_loopnr_4	150087
result2_0_3_2_3_0_1_1_0_1_0_1_1_loopnr_5	150074

Table 23: Results number of records for compression – DISSOLVE

INTERSECT in default settings in ArcGIS on the fat client, Windows 7

After running the INTERSECT real data workload in its default settings, the log data revealed that two of the 5 runs resulted in a different number of records (39 records difference) and also a higher execution time, which is a good warning to be careful with medium values of the execution time. This pattern has not been recorded during other INTERSECT workload, which is the synthetic workload and the workload on the big data computer.

source_tbl_or_fc	Number of records	Execution time
result4_0_3_2_3_0_1_0_0_1_0_0_1_loopnr_1	85488	298
result4_0_3_2_3_0_1_0_0_1_0_0_1_loopnr_2	85488	296
result4_0_3_2_3_0_1_0_0_1_0_0_1_loopnr_3	85527	362
result4_0_3_2_3_0_1_0_0_1_0_0_1_loopnr_4	85527	351
result4_0_3_2_3_0_1_0_0_1_0_0_1_loopnr_5	85488	295

Table 24: INTERSECT default - differences in number of records and execution

6.8 Conclusions benchmark

After analysis, visualization and validation of the results, the following results can be stated:

6.8.1 ArcGIS Desktop 10.1 shows partial improvement, in most tools, but substantial degradation performance of JOIN and DISSOLVE

It is certainly not a "black or white" difference in performance, as perhaps perceived at Spatial Statistics, with a good performance of Workstation versus bad performance of Desktop. UNION, INTERSECT, NEAR, SUMMARY STATISTICS and FREQUENCY even perform better in Desktop 10.1. However, the DISSOLVE and JOIN show considerable loss in performance, most probably because of the difference in data structure.

6.8.2 Limits of Arc/ Info Workstation in scalability

Based on the benchmark results, ArcInfo Workstation cannot be viewed as the most preferable environment: The performance is not better on all fronts, only for the DISSOLVE and the JOIN. Additionally, scalability is very limited: Creation of larger input datasets than 1.000.000 records in Workstation often leads to a crash of the Arctool due to its size. The use of ArcInfo Workstation clearly shows better performance in the DISSOLVE with real data and input datasets from 10.000 to 1.000.000 records synthetic data.

6.8.3 Development of scalability with synthetic data linear, except for INTERSECT and UNION

The geoprocessing tools have been tested on scalability and algorithm with the use of simple synthetic data. While DISSOLVE and NEAR showed clear linear results, INTERSECT has shown a faster execution time during the first step of increasing the input data, but showed a linear curve up to the use of a combination of 150.000.100 and 10.000.000 record dataset, whereas the combination with 50.000.000 records as a second dataset showed a sharp upward curve. The values of the UNION tests remained almost stable. The reason for these results could be the large different in number of records between the first and the second input dataset.

6.8.4 Exclusion of noise factors network and use of workspace does not lead to performance improvement

No influence could be detected of network activity because of access to a user profile, data ore a script. Storage of in- and output data in same or different workspace did not make a difference either.

6.8.5 Chosen optimization strategies in ArcGIS 10.1 do not show improvement

Sorting and indexing, as well as compacting and compressing do not show any improvement in performance with the used tools (UNION, INTERSECT, NEAR and DISSOLVE) in ArcGIS 10.1, compared to the omission of these methods. Compressing even shows a much slower execution time than the scenario where no compression is applied. The benchmark showed a negative impact on performance with compressed data and a small difference in output data.

6.8.6 Upgrading hardware does not provide satisfactory results

The display of the benchmark results in Figure 46 showed no improvement based on hardware (or hardware combined with ArcGIS 10.2.2). The tool where the big data configuration showed improvement was the NEAR, which has been redesigned in ArcGIS 10.2.1 and therefore shows large improvement in performance. Hardware upgrading could improve performance of tools that are already "organized more efficiently".

6.8.7 Administrative processing results show that performance issues are not purely related to geoprocessing and spatial data structures

Whereas FREQUENCY and SUMMARY STATISTICS perform better in ArcGIS desktop 10.1, a problem remains with the JOIN FIELD, which shows a significant deterioration in performance compared to the JOIN ITEM of ArcInfo Workstation. This has been indicated in Figure 46 in section 6.6. Workarounds using the ADD JOIN have been proposed by ESRI, but showed unstable results in ArcGIS Desktop and ArcGIS Pro. Surprisingly, there is no improvement of the tools in both big data configurations.

6.8.8 Validation of output data shows irregularities after compressing the data and during the INTERSECT (default)

Out of 5 test runs, the default scenario for INTERSECT, showed 2 runs with output datasets containing higher number of records (both with the same number of records) and related higher execution time. The compression of input datasets resulted in deviating output datasets for UNION and DISSOLVE. In both cases, it is worrying that these differences in output occur during the runs of the same scenario, without change of parameters. Certainly, this has to be researched further by ESRI and Statistics Netherlands.

Chapter 7: Lessons Learned From Similar Organizations

The goal of consulting other organizations is not to collect quantitative data on use and performance of geoprocessing tools, but to provide some guidance and lessons learned to help during the decision making process. Different channels have been used to get in contact with similar organizations: The LinkedIn page of the AGGN (Dutch ArcGIS user group), the ESRI ArcInfo Workstation Forum and the network of Spatial Statistics staff via INSPIRE working groups. Via the Workstation Forum, only a representative of the USGS has reacted[8], via the Spatial Statistics network, initially 6 have reacted. These 6 contacts have received questionnaires. Finally, the questionnaire resulted in 3 reactions. The amount of detail provided in the questionnaire varied per respondent. Statistics Portugal also provided additional documentation. Additionally, an interview has been held with the PBL (Netherlands Environmental Assessment Agency), which allowed for more interaction than with the other organizations, which received questionnaires.

The domains of the statistical agencies are of course different from the USGS and the PBL: The core business of the USGS is scientific research on natural resources and climate whereas the PBL deals with policy research in the fields of environment, nature and spatial planning. In these organizations, geo-information is part of the core business. This is also reflected in the number of GIS users, shown in Table 25. The position of geo-information also varies per organization: The respondents of Statistics Portugal, Statistics Italy and PBL work for a separate organizational unit that deals with geo-information, whereas Spatial Statistics is one of the statistics producing teams within the sector Environmental, Energy and Spatial Statistics.

As private sector companies are often involved in support of migration projects, an interview has been held with Aris (Desabandu et al., 2014), a company that supports public agencies, research institutions, municipalities and non-profit organizations with the use of ESRI, Oracle and open source products. They have experience with support of migration from ArcInfo Workstation to ArcGIS Desktop and have supported the PBL in their migration. The results of this interview will be included in the different sections of this chapter, although not in the table, such as the other organizations.

The results of the interview and the questionnaires (for Statistics Portugal a presentation for INSPIRE has been added as well) are presented in the following sections, covering the GIS users, geoprocessing activities, performance optimization, ICT infrastructure, migration experience and lessons learned. The results of the questionnaires have been stored in tables as well, available in Appendix 5. In each table, the information has been presented for Statistics Netherlands as well (except for sections “Migration” and “lessons learned”, to provide a quick way to compare their situation with the respondents.

7.1 GIS users

The goal of this section is to get an impression of the number of users and of the currently used GIS software. With light users, occasional simple visualization, editing and simple analysis are involved, whereas heavy users regularly apply heavy analysis (for example a series of tools via Modelbuilder or scripting).Table 25 in the Appendix shows that only Statistics Netherlands and Statistics Portugal have approximately the same number of heavy and light users. The total USGS number of GIS users is not known, but the percentage of heavy users is much higher than of the light users, contrary to the other listed organizations. It is, regarding domain and scale, least compAppendix 5.arable with the other organization.

Surprisingly, only Statistics Portugal has a spatial database (Oracle), although the number of GIS users is not that high. They have also combined it with the ArcGIS 10.2.2 version. The Netherlands Environmental Assessment Agency used to have Oracle but discarded it because maintenance of the data required too much knowledge and workforce to tune and maintain the database and has been less flexible. This may have been one of the reasons that the use of the database did not lead to better performance. On the other hand,

[8] Of the USGS South Dakota Water Science Center in Rapid City

Statistics Portugal has introduced ArcSDE around 2003-2005 out of a need for better data management and data dissemination. Overall, use of open source software or other proprietary GIS software is very limited.

7.2 Geoprocessing and performance optimization

This section covers the geoprocessing workload of the organizations and the current performance optimization methods. The type of analysis that Statistics Netherlands is doing has most in common with Statistics Italy and Portugal, which is not surprising, given the fact that they belong to the same type of organization. Spatial Statistics shows a higher variety of products, but the amount of information available on their products is also the highest. The PBL works with environmental models that also include spatial-temporal data which leads to a higher data volume. At the USGS (Water Science Center), mostly raster analysis is used and datasets are prepared for hydrographic analysis.

The geoprocessing tools are mostly executed by script and/or modelbuilder. Partitioning and tiling is mostly used to improve performance. The Netherlands Environmental Assessment Agency also uses multiprocessing (via scripting), although it has been judged as useful only for CPU bound tools and for calculations where there the total execution time (including pre- and post-processing) is still faster than the linear mode. The ArcGIS servers in the data centre of the PBL can be used as a grid for multiprocessing. Optimization by adaption of the index or sorting, or the application of compression or compaction is not mentioned, although this has also not been mentioned as an example in the questionnaire. There is not much time available to experiment with performance optimization, only at the PBL and USGS.

7.3 ICT Infrastructure

The questions within this section were asked to find out if the needs for the GIS users are met by the ICT departments of these organizations and how these needs are communicated. The communication of ICT needs of the GIS users seems to be organized very differently at the listed organizations, although the level of detail in the provided answer is very different per respondent. It also seems that at organizations such as the Netherlands Environmental Assessment Agency and USGS, with a large GIS user base and a more research related environment have more ICT support and facilities at their disposal. The use of virtual desktops at the Netherlands Environmental Assessment Agency is very interesting, since virtual desktops have not worked well for Spatial Statistics. Statistics Italy and Portugal use local processing, like Statistics Netherlands.

7.4 Migration

The other organizations migrated in the early 2000's to a higher Desktop version (or ArcSDE). Although some ArcInfo Workstation or ArcView users remain, they do not take responsibility for complete production processes like at SN. Of course, it cannot be ruled out that the other organizations migrated with a "smaller" workload for example smaller datasets and different calculations. The Environmental Agency has been supported by the company Aris during its migration, others like Statistics Portugal and Italy have been initially supported by ESRI. Aris has described different motivations to migrate from AML (as programming language of ArcInfo Workstation) to Python (ArcPy as library for ArcGIS Desktop):

- The use of low level programmed tooling based on ArcObjects (development environment using Visual Basic for Applications, Java or C#), which allows for more custom made tooling, and as possibility to improve performance
- Dealing with the data more efficiently
- More advanced hardware resources, but not as a sustainable solution

Aris has experienced during migration projects that some tools are faster in AML, but overall performance of the new environment has not been slower. Moreover, correct output of the geoprocessing tools is most important (Desabandu et al., 2014).

The experiences with the performance of the new software have not been completely positive or negative mostly: the Environmental Assessment Agency discarded a number of models because of instability of ArcPy; other experiences have been mixed up to positive.

7.5 Lessons learned

A number of aspects have to be kept in mind reading the questionnaire/interview results: The level of detail in answering the question and English writing skills are very different per filled form. It is not known if those interviews would have been answered differently with a differently formulated question. After analysis of the provided information it can be concluded that, regarding users, workload and facilities, the situation of at Statistics Netherlands can largely be compared with the situation at Statistics Italy and Statistics Portugal: They are conducting similar analyses and are similar in number of users. Like Statistics Netherlands, time is very limited to work on performance optimization. The questionnaire contained a question on lessons learned, but the lessons learned that are stated below are the result of the total content of the questionnaires:

1. The transition has been a gradual process for the respondent organizations. The respondent of the USGS has called the migration a gradual process instead of a migration project, but also the answers of other organizations show that a few ArcInfo Workstation users still remain, even after such a long time, app. 10-15 years. An important difference with the situation at Statistics Netherlands is that they started to migrate a long time ago, when ArcInfo Workstation was not at the end of its product cycle.
2. The influence on ICT infrastructure decision making in terms of facilities for geoprocessing will depend on the user base and the organization of the GIS users: are they represented in a central, supporting Geo-information unit, or does a unit from “the business” like Spatial Statistics represent their needs to a certain degree?
3. The PBL shows interesting options in optimization, but also has a more flexible (Geo) ICT infrastructure.
4. Open Source or other proprietary software is largely new territory and not really investigated on functionality and performance.
5. For most of the Statistical bureaus and the PBL some part of the migration has been outsourced, at least the initial phase of the migration.
6. The evaluation of the migration to a higher ArcGIS version is mixed and varies per tool or model. It is also dependent on the user needs of the organization: Does performance have a high priority or rather maintainability, organization of the data?

Chapter 8: Trends in geoprocessing

During the past years, “big data” have emerged as a consequence of new data sources. The technologies that are described in this chapter are being researched for a number of years already, but not yet regularly applied for a number of reasons, for example lack of knowledge. Often, the solution is a combination of technologies: distributed storage via grid computing, for example, is not possible without an efficient organization of the data with the use of, for example, space filling curves. Grid computing is used to facilitate parallel processing.

The following section will describe a number of technological geoprocessing trends: section 8.1 will describe in short new forms of data collection that lead to new data formats, whereas 8.2 will deal with cloud services. Section 8.3 will describe the MapReduce principle and 8.4 will cover the application of GPU processing. In section 8.5, these trends will be discussed in regarding their applicability for Statistics Netherlands.

8.1 New forms of data collection and data formats

New ways of capturing spatial data push the need for new ways to store and disseminate that data. Lidar (light detection and ranging), for example, is a remote-sensing technique that “uses laser light to densely sample the surface of the earth, producing highly accurate x,y,z measurements” (ESRI, 2014e). The Lidar measurements are stored in a massive point cloud data format, storing x, y z and possibly other attributes. Point cloud data are stored in file-based systems or relational database systems (van Oosterom et al., 2015). A well-known file-based public point cloud data format is the binary format LAS. It is mostly used for LIDAR point cloud data, but can also be applied for exchange of other 3-dimensional data. It also maintains the LIDAR specific information of the data (ASPRS, 2012).

The LAS format also has a compressed version, called LAZ (van Oosterom et al., 2015). ESRI has developed a compressed point cloud format as well, the ZLAS format, which is useable with ESRI GIS software (van Oosterom et al., 2015). Possibilities for a Point Cloud Spatial Database Management System rare being researched (van Oosterom et al., 2015). At the moment, no explicitly stated need exists for the application of point cloud data, but it might change in the future.

8.2 Cloud services

Providing on premise hardware facilities for big data geoprocessing can lead to very high costs. Therefore, more organizations have started to research the possibilities of cloud solutions. Cloud environments offer services for data storage, application or infrastructure (Yue, 2012). There are different cloud environments such as Google App Engine (GAE), Microsoft Azure or Amazon Elastic Compute Cloud (EC2). There are different forms of cloud computing services (Kouyoumjian, 2010):

- **Software as a Service (SaaS):** The services offers end-user applications, for example an off-premises asset management system.
- **Platform as a Service (PaaS):** This service does not provide a ready-to-use application but an application platform or middleware as a service to create custom applications. The Windows Azure platform of Microsoft is an example of a PaaS
- **Infrastructure as a Service (IaaS)** delivers an off premise infrastructure of storage, operating systems or computing power. Amazone Elastic Compute Cloud (EC2)ArcGIS Server is an example of such a service.

Besides the apparent advantages of cloud environments, also challenges of cloud environments can be noted:

- Information security: Statistics Netherlands often uses micro datasets that contain privacy sensitive information. Therefore, the ICT infrastructure of Statistics Netherlands is almost completely closed (see also chapter 3).
- Interoperability, because of the diversity of data formats and geoprocessing platforms (Yue et al, 2012). Adhering to OGC standards.

8.3 Hadoop and MapReduce

Hadoop is an open source implementation out of Google's framework MapReduce, built for distributed storage and process large amounts of data in parallel via a distributed infrastructure of computer clusters (Hoel & Park, 2013). The raw data on Hadoop is not yet suitable for geoprocessing, structuring the data is therefore essential. Structuring of data in Hadoop employs 2 main techniques: sorting and splitting the data into shards. These shards store the data that is spatially related on the same machine (Hoel & Park, 2013).

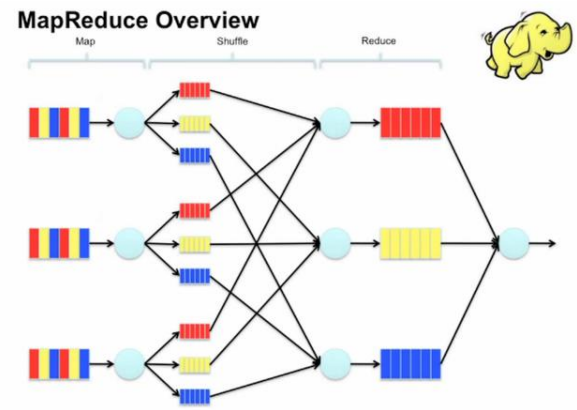


Figure 49: MapReduce principle (Janakiram, 2012)

As of late, ESRI has also implemented GIS tools for Hadoop (http://Esri.github.io/gis-tools-for-hadoop/). The Port of Rotterdam Authority has used the combination of Hadoop and ArcGIS to map ship movements for a certain period (Desabandu & Eijkelenboom, 2014).Even if all these solutions that have been discussed lead to considerable performance improvement, the cost of implementing these solutions will be very high: It requires resources to build knowledge of these methods.

8.4 GPU enabled processing: CUDA

Graphical Processing Units (GPU) have emerged as a comparably affordable technology to be used for "General Purpose Computing"[9], while other high performance computing resources (for example parallel computing resources) have been available only for select research groups (Zhang & You, 2012). For optimal use of the GPU technology, data structures and algorithms are needed that are suitable for parallelization (Zhang & You, 2012). Many GPU's are enabled with CUDA, a parallel computing platform and programming model (NVDIA, 2015). Some proprietary GIS software applies GPU processing, for example Manifold (Manifold). ESRI seems to have implemented GPU computing primarily for visualization (Desabandu & Eijkelenboom, 2014), although it has been tested already in 2010 (ESRI, 2010a).

[9] General Purpose computing on Graphics Processing Units (GPGPU)

8.5 Summary: Consequences for Statistics Netherlands

At the moment, Statistics Netherlands uses large datasets, but in the “traditional” 2 dimensional data. In future, there would probably be less need for 3 dimensional data like the point cloud data, but more spatio-temporal data, which can be time sequences of a certain area, telephone or sensor data with a location. Team Methodology is already experimenting with R Spatial to analyse and visualize traffic sensor data. For implementation of cloud services at Statistics Netherlands, only geoprocessing for open data would be viable. Furthermore, additional research would be needed to assess which cloud configuration would be most suitable. Such an option could be investigated with the help of the innovation lab. It is important, however, to have an overview of the total costs that would be involved in setting up the infrastructure. The Hadoop/MapReduce framework as well as GPU enabled processing need efficient structuring of the data and knowledge of those frameworks. It is difficult to assess whether the cost of investing in data preparation and knowledge building will be outweighed by the improvement in performance. Even if the described trends are not implementable on the short term, future developments should be tracked by Spatial Statistics in cooperation with related teams and proofs of concept could be organized to assess their usefulness for the organization.

Chapter 9: Discussion, conclusions and recommendations

This chapter starts with a discussion of the research process (9.1), followed by a the answers of the subquestions. These findings will be subject to discussion before answering the central research question: *Which alternatives to the current geo-ict infrastructure can be proposed for Statistics Netherlands that meet performance requirements of its geoprocessing activities and are suitable for implementation within the organizational constraints of Statistics Netherlands?* The final conclusion will be followed by recommendations (9.2).

9.1 Discussion and conclusions

This research project has provided an interesting view on geo-information processing at a governmental agency. The production of these data is initiated by law, governmental policy and societal needs and therefore has to apply high quality standards. Therefore, production and quality requirements play an important role. They leave no room to experiment in the production environment, which means that more resources have to be reserved for research on new technologies to produce current products more efficiently and to produce new products. Many governmental agencies are therefore moving forward to an up to date (geo) ICT infrastructures at a slower pace. At the same time, the amount of data is growing fast and the academic world and private sector work on solutions for the growth in data. Methods to deal with the large amount of data like MapReduce and Grid Computing are important developments but require logical organization of the data, high investment in expertise and hardware. The same can be said for new data formats that have evolved out of new data capture methods.

For the research, a broad approach has been chosen, taking into account the technical aspects of geoprocessing performance but also the organizational aspects that come into play when changes in the technical infrastructure are needed. These organizational aspects have been covered in (part of) chapter 3 and 5. The information has been gathered by interviews with representatives of the Spatial Statistics staff, a staff member of the IT department and the head of the innovation laboratory. They all have been essential to answer the first research question: *Which technical as well as organizational bottlenecks can be identified during geo-information processes at Statistics Netherlands and how are they currently dealt with?*

Still, a more complete analysis could have been made with a GIS user from another department, e.g. team Process Development and Methodology that are R Spatial users or another ArcGIS user. Now it is not so clear what the needs of users outside the team Spatial Statistics could be. Furthermore, the technical body of work included the literature research on benchmarking, the benchmark development, benchmark execution and the interpretation of the results. The benchmark could not be exhaustive in its choice of optimization possibilities, metrics and validation. The recommendations by ESRI (Iparraguirre, 2014) which have been directed at improving the fat client infrastructure at Spatial Statistics have not been implemented yet within the infrastructure of Spatial Statistics. The response of the contacted organizations for research question 6 has been relatively low, although valuable information was provided. It is possible to deduce some lessons learned from the given answers. Best practices are harder to discover because the organizations vary in a number of aspects: size, products or number of GIS users and much. A lot of variation could be observed in the level of detail of the answers in the questionnaires.

This research has been new in different ways: Proprietary GIS software has been tested within a real life situation, using both real life and synthetic data. Additionally, state-of-the-art computer resources and the newest ArcGIS version ArcGIS Pro have been tested. This technical aspect of performance benchmark has been combined with documentation and interview reports that mainly deals with the organizational aspect of performance optimization. This is necessary to be able to reach the goal stated in the title of the thesis: “Towards a sustainable geoprocessing environment at Statistics Netherlands through performance benchmarking”. Which sustainable alternative can be advised to them? First, all research questions will be answered before in the subsections 9.2.1 through 9.2.6 .

9.1.1 Which technical as well as organizational bottlenecks can be identified during geo-information processes at Statistics Netherlands and how are they currently dealt with?

Diversity of users, applications and performance needs

Researching technical documentation and speaking to the members of team Spatial Statistics showed different types of users within and outside the team Spatial Statistics. Spatial Statistics provides a small number of heavy users, skilled in ArcInfo Workstation and AML and partially in ArcGIS Desktop and Python. Parallel to these users, their colleagues at the department of Process Development and Methodology, use R Spatial for statistical analysis on spatial data. There is no real exchange in knowledge between Spatial Statistics and Process Development and Methodology. The light users apply different software via the virtual desktop, such as ArcView, ArcGIS Desktop and PX Map. This situation makes is difficult to support and optimize the use of spatial data from the point-of-view of the users and the IT department.

User experiences with ArcGIS Desktop

A number of users have experienced instabilities with the application, for example crashing of the application after running a tool (no specific tool). The role of the user profile is not really clear: making a new user folder on %appdata% helped. For some tools the desktop application showed a much slower execution time for some users, e.g. the (administrative) JOIN, but also UNION. Al lot of methods to optimize performance and forestall instabilities have been applied: Dataset partition (on attribute or record level), local processing and maintaining the AML /Workstation tools, but there is no sustainable strategy available.

Lack of means of IT department and Spatial Statistics for application management

There are only 2 staff members of the IT department who are doing tasks such as licensing, troubleshooting for the GIS users at Spatial Statistics or even Statistics Netherlands. At the same time, the production processes and available resources at the spatial team leave very limited time to fundamentally think about non-functional requirements such as performance, set up SLA's (Service Level Agreements) and look for means to fulfil the non-functional requirements together with the IT department.

9.1.2 Which (technical) performance factors can be identified for geoprocessing based on literature examples and on experiences at Statistics Netherlands and how can these factors be prioritized and evaluated with the design of a benchmark, based on the performance requirements of Spatial Statistics?

Research to answer this research question showed that the foremost performance requirements are: reliable results, stability and execution times that are not slower than the current ones in ArcInfo Workstation. Related work has shown that the development of performance benchmarks for geoprocessing environments has been developing slowly, while it is often situated within an academic environment with highly configurable test beds. The software that has been used in these benchmarks have mostly been spatial databases. This benchmark is situated within a production environment, using only one software vendor and file-based systems. This automatically leads to a reduction of parameters of benchmark factors. Therefore it is important to grasp that the benchmark will only provide information on performance in ArcGIS Desktop 10.1, 10.2 and ArcGIS Pro, which are file based systems. The first steps of the benchmark are baseline tests which use the settings in ArcInfo Workstation 10 and ArcGIS Desktop 10.1 and scalability tests with synthetic data to assess the efficiency of the algorithm. Based on literature study, experiences at Statistics Netherlands and interviews with ESRI a number of factors have been identified and categorized in three groups within a benchmark:

- Factors where the role of network or workspace interference is assessed
- Factors that assess the role of performance optimization in ArcGIS 10.1: spatial access methods like sorting and spatial indexing in ArcGIS 10.1, compacting and compressing
- Factors that assess the role of hardware configuration and a new 64 bits desktop system, ArcGIS Pro.

9.1.3 What are the results and conclusions of the performance benchmark tests?

The benchmark results indicate that the long term investment in application and algorithm design yields the best results in performance. The leading example for this conclusion is the (in ArcGIS 10.2.1) improved NEAR algorithm which resulted in substantial improvement of performance, whereas DISSOLVE could never reach the performance level of ArcInfo Workstation, most probably because of the controlled topological data structure of the coverage.

9.1.4 What are geo-ict related trends (e.g. role of performance in following releases of ArcGIS or other proprietary or open source software) that could influence the recommended migration scenario for Statistics Netherlands?

The presented trends are: Cloud Computing, Hadoop/MapReduce and GPU driven geoprocessing. The presented trends would probably not influence the migration scenario or the production environment for the short term. It is more realistic to apply the presented technologies for innovative projects in cooperation with the innovation laboratory, such as for example a cloud solution that uses only open data.

9.1.5 What are lessons learned from organizations with a profile similar to Statistics Netherlands that process large spatial datasets and can provide information on performance evaluation and its role within decision making related to geo-ict infrastructure?

The situation at Statistics Netherlands can be compared with the situation at Statistics Italy and Statistics Portugal regarding types of analyses and number of users. The time is very limited to work on performance optimization. Yet, Spatial Statistics has a position within the business of Statistics Netherlands and is mostly a production unit and only for a small part a supporting unit on Geo-Information. The size and position of Spatial Statistics makes it difficult to acquire suitable resources and to experiment with methods to optimize performance. The lessons learned that have been obtained by the interview and questionnaires can be formulated in the following points:

1. The transition has been a gradual process for the respondent organizations, but for Statistics Netherlands the urgency to migrate is higher because of the high production load in AML and the product cycle of ArcInfo Workstation.
2. The influence on ICT infrastructure decision making in terms of facilities for geoprocessing will depend on the users of geo-information and their organization:
3. The PBL shows interesting options in optimization, but also has a more flexible (Geo) ICT infrastructure.
4. Open Source or other proprietary software is largely new territory and not really investigated on functionality and performance.
5. For most the Statistical agencies and the PBL some part of the migration has been outsourced, at least the initial phase of the migration.
6. The evaluation of the migration to a higher ArcGIS version is mixed and varies per tool or model. It also depends on the user needs of the organization.

9.1.6 Which (geo) ICT scenario is most suitable for the organization including implementation steps and quick-wins?

Deciding on a migration scenario for Spatial Statistics is difficult, because of its high impact on working processes and resources, even more because Spatial Statistics has been building so much expertise and products in ArcInfo Workstation and AML and has been innovative in developing these products at an earlier stage. Without the expertise in ArcInfo Workstation, the current number of products would probably not be that high. The success of these products has also led to a higher number of processes, developed in AML scripts that have to be reengineered. Continuing with AML/Workstation would be the easiest option on the short term, but not sustainable: support stops at the end of this year and ArcGIS Pro, which will be the successor of the Desktop product line, will not support the use of coverage tools. Partitioning of datasets will continue to be needed because of the size limits of the coverage.

At the same time, an interesting development is taking place at departments like the methodology department and the innovation laboratory: They are starting to use large spatial datasets and use open source software. This leads to scattered Geo ICT knowledge and production processes at Statistics Netherlands. Therefore, it would be advisable to cooperate on projects with the use of spatial data and to invest time to exchange expertise. Staff members that are very knowledgeable with AML, have to reach the same standard in Python. This will also cost time, although some support provided from other teams with Python experts who are not experienced with Geographic Information Systems is considered. Therefore, some training in GIS will be necessary for the Python experts and further training in Python will be needed for the Spatial Statistics Staff members. Spatial Statistics has already started to provide Python training to a number of staff members.

The benchmark results indicate that the long term investment in application and algorithm design as well as efficient data structures yield the best results in performance. The leading example for this conclusion is the (in ArcGIS 10.2.1) improved NEAR algorithm and the weak performance of the DISSOLVE in Desktop versus ArcInfo Workstation. This also indicates that other software solutions using different algorithms or a data structure that organizes its data more efficiently would be interesting to investigate further.

However, the goal is also to achieve a sustainable geo-ict infrastructure. The term “sustainable” is especially important because it entails more than satisfactory geoprocessing performance but also systems stability and reliable results. Additionally, with the fast developments in GIS technology that push the product cycles of software developers like Esri, it is legitimate to state that a currently sustainable solution will not be sustainable anymore in approximately 5-8 years. Therefore, the next migration step should be the first one of a development that should be supported by the four “wares” of Information Systems:

- Software: reengineering of current AML scripts, acquirement of software that adheres to functional and non-functional requirements.
- Hardware: suitable hardware and network bandwidth
- Humanware: knowledge and skills in new technology, application management
- Orgware: reengineering of processes to meet production deadlines

These findings lead towards an answer for the main research question:

Which alternatives to the current geo-ict infrastructure can be proposed for Statistics Netherlands that meet performance requirements of its geoprocessing activities and are suitable for implementation within the organizational constraints of Statistics Netherlands?

There is not one correct solution, different alternatives can be considered. However, considering the previous thought on the lifespan of sustainability, the first alternative should be considered:

Alternative 1: Keep small scale infrastructure within ArcGIS Desktop

Spatial Statistics stays within the ArcGIS Desktop suite but should consider quicker migration to 10.2 or rather 10.3 because more performance problems are resolved within these versions. According to the benchmark results, at least the NEAR has been improved in ArcGIS Desktop 10.2.1 and according to ESRI, several bottlenecks have been addressed in 10.2 and 10.3. Additionally, performance problems will resolved mostly just before system upgrades. An advantage of 10.3 is the availability of ArcGIS Pro, which should lead to further testing of that application. The possibilities to optimize ArcGIS Desktop 10. 1 remain very limited. This will mean that Spatial Statistics will need less expertise in-house on performance optimization, but will remain dependent on ESRI product development processes. This dependency could be countered by cooperating more closely with other ArcGIS users nationally and internationally to press for the needed improvements.

Consequence of this option would be the redesign of AML scripts to Python 2.7.x and 3.4 (the version that is supported by ArcGIS Pro), possibly also a change in methodology per script. Several disadvantages are part of

this solution: some tools would be improved in performance, such e.g. the NEAR, but others like the DISSOLVE or the JOIN FIELD would still be a problem for performance. It would be advisable to assess per script, which tools could form a bottleneck and if they could be replaced, executed within a different environment or if the organization or the production process has to be changed. For example, planning certain processes earlier could be advisable if the performance problems remain unsolvable within the Desktop environment.For the long term, upscaling of the Geo ICT infrastructure is recommendable, based on the current situation of Statistics Netherlands and lessons learned of other organizations:

Alternative 2: Invest in an up-to-date infrastructure with other departments

To be more in control of performance optimization means that a more configurable Geo ICT infrastructure is needed. Spatial Statistics is too small to make such an investment in infrastructure and needed expertise ("humanware"). Therefore it should cooperate with other departments at SN who have a need for spatial data analysis. Such a (geo) ict environment could include several components: a spatial database such as Oracle combined with ArcSDE, an open source spatial database such as PostGIS or a file based environment. Additionally, extension of the benchmark towards spatial databases would be needed. This need is especially apparent if the volume of data is expected to grow in future. This solution would require even more resources and also a different embedding of geo-information within the organization. A possibility could be the formation of a "virtual" geo-information team. Such a step requires additional research and start-ups of Proofs of Concept (POC). Therefore, The third alternative should be used as a complementary step.

Alternative3: Organize innovative projects with internal and external partners

In addition to the more short-term solution Spatial Statistics should start a number of projects dedicated to the application of new geo-ict technologies. For example, a project involving Spatial Statistics and other teams in cooperation with the innovation laboratory. It is recommended to use this alternative to prepare the step towards Alternative 2, but also as a complementary step to Alternative 1.

The recommendations have been divided into recommendations to improve the performance benchmark and recommendations for further steps that Statistics Netherlands could take.

9.2 Recommendations

9.2.1 Further migration steps

Establish geo-information vision for the long term

Without a perspective of the goals that have to be achieved, it is impossible to define what is necessary to achieve that goal. For the long (or even mid-) term, Statistics Netherlands has to form a common vision and goal regarding the role of geo-information. This is a process that has been started already, partially because of influence from the director general of Statistics Netherlands, who has asked all teams to formulate the purposes that the team serves. It is clear that geo-information is used in more teams than Spatial Statistics. At the moment, this is happening independently in each team. By operating from a common vision and being open to information needs of other teams and departments, more impact and influence can be exerted within the organization. Important partners could be the methodology and innovation department, but also other teams, like real estate or environment

Collect information on use of R Spatial

Use of R Spatial is out of scope for this research, but should be part of establishing the long term vision for the use of geo-information. There is no information available on products made with this software, even less on types of analysis and their performance.

Try to cooperate with other ArcGIS user organizations to lobby for improvements of certain tools

The larger the base of users that demands redesign of tools that do not perform well such as the JOIN and the DISSOLVE, the higher the chance that ESRI will put more effort into performance improvement (of these tools). This has been proved by the redesign of the NEAR and a number of raster tools for ArcGIS Pro.

Define Service Level Agreements (SLA's)
Currently, the non-functional requirements are not explicitly stated: the performance requirements have been limited to the current performance with ArcInfo Workstation. Defining SLA's will aid communication with the IT department.

9.2.2 Steps to extend the benchmark

Repeat synthetic data "stress tests" on Big Data computer
The boundaries of the prioritized tools have been tested within the fat client environment. Due to the limited available time, the synthetic data test series have not been tested and, if possible, extended on the big data computer, for example, a DISSOLVE on a synthetic dataset up to 100.000.000 records.

Repeat real life data workload on fat client with ArcGIS Pro
The impact of ArcGIS Pro could not be clearly evaluated, because it has been tested within a different ICT infrastructure, not connected with the Statistics Netherlands network and with the availability of better hardware resources, such as the SSD disk, higher memory and higher CPU capacity. In the third quarter of 2015, ArcGIS Pro will probably be installed on a fat client at Spatial Statistics. Then it will be possible to judge whether the further improvement of the NEAR has occurred due to the hardware capacity or due further optimization of the NEAR algorithm in ArcGIS Pro.

Repeat workload scenarios at the Netherlands Environmental Assessment Agency
Statistics Netherlands regularly communicates with the Environmental Assessment Agency on data deliveries or exchange s of experiences. With a much larger group of GIS users and its more research driven working processes, the agency has been able to experiment more with optimization of performance and to receive more resources for its ICT infrastructure. Conducting geoprocessing via a virtual desktop has been successfully implemented although ArcGIS Desktop 10.1 is used. It would be very interesting to run (part of) the benchmark at the Environmental Assessment Agency and to compare the results.

Repeat INTERSECT default workload on the fly
The results of the INTERSECT workload using the default settings have resulted in different output tables: out of the five runs (executed via script), two show a difference in number of records. A possible reason could be a memory problem due to the use of iterations. Therefore, it is recommended to repeat the tests with five runs, but not within a script but on the fly via ArcMap. If this still leads to the same results, ESRI should conduct further research into the matter.

9.2.3 Scientific Outlook

Some of the ideas for further research could be an extension of the current benchmark but require more than just minor adaptions, especially if different GIS applications are involved or different types of data. Not only the technical but also the organizational perspective could provide opportunities to set-up a follow-up research project, for example successful reengineering and implementation of information processes.

Extend performance benchmark by comparing different file based proprietary and open GIS
Would it be possible to execute a workload similar to this benchmark in other proprietary (e.g. Manifold, Geomedia) or open (QGIS) software? Combined with a survey or interviews on performance and scalability of these applications, the benchmark could be redesigned. This will require time and support of experts, because of the difference in geoprocessing functionality.

Extend performance benchmark with spatio-temporal data

Currently, Spatial Statistics has not used Spatio-temporal data for statistical products, whereas team Process Development and Methodology has built some experience with these data. It could be a good opportunity to cooperate on a project that involves performance analysis of geoprocessing of streaming spatial data in R Spatial, Arc GIS and another alternative, for example QGIS.

Find suitable strategies to conduct successful reengineering of processes based on performance benchmark results?

It would be very interesting to continue where we left off with this research project: In the conclusions it has been established that some tools will form performance bottlenecks in ArcGIS Desktop 10.2 or higher. Are there proven strategies to adapt processes based on performance benchmark results?

Appendix 1: Description of spatial products Statistics Netherlands

This appendix describes in short the most important products of team Spatial Statistics, which have been referred to in chapter 1. These have been regular products for many years already. New products are also being developed at the moment, but their status is not mature yet.

Bestand Bodemgebruik (BBG – Land use File for the Netherlands)

The aerial imagery of the Netherlands is interpreted and encoded with land use codes using datasets like TOP 25 and Top10 NL (BRT) as well as Locatus (a database containing consumer locations like shops, hotels, restaurants, etc.), Woning register (register of apartments) and Bedrijvenregister (business register). The BBG is updated every 2 years, northern and southern half of the country are alternately published every year. This product is more a results of data entry than geoprocessing: The used datasets are used as a reference to enter the correct land use code. The input datasets that are used are:

- BRT/Top10NL
- Aerial imagery (as reference layer)
- BBG previous version (since introduction new production method)
- Locatus (reference layer)
- Woning register (reference layer)

Financiële Verhoudingswet

Municipalities receive subsidy from the Ministry of the Interior and Kingdom Relationships according to different criteria. Some of these criteria are spatial, such as the total area of built-up areas, rural areas, soil quality and land-water rate. These values are calculated by the Spatial Statistics team. Large sums of subsidy are assigned based on these calculations, therefore correct input data, calculation method and interpretation are essential. The main input for this product is the BRT, which is converted back to Top10Vector format in coverage. The BRT is delivered processed per map sheet.

The input datasets that are used are:

- BRT (Top10NL)
- GBR (geografisch Basisregister)

Proximity Statistics

Two times a year, proximity of facilities to residents via the road are calculated per district/neighbourhood. The facilities vary from hospitals to schools, or day-care institutions or libraries. This information is very helpful in providing location intelligence for municipalities and businesses. The input datasets that are used are:

- BAG (Key Register of Addresses and Buildings)
- NWB (Dutch Road Dataset)
- Datasets of facilities, e.g. Locatus, or facilities per sector (e.g. day care facilities, libraries)

Wijk- en Buurtkaart – Municipality- District- and Neighbourhood statistics

This product contains the geometry of municipality-, district-, neighbourhood-borders combined with aggregated statistics. These data (3 datasets) are produced at the beginning of each year. The lead time for statistics calculation is different per statistic type. The geometry is derived from the Key Register Kadaster, CBS Land Use Dataset and municipalities, the statistics from the “Kerncijfers Wijken en Buurten” (key figures of districts and neighbourhoods). The input datasets that are used are:

- “Bestand burgerlijke gemeentegrenzen” derived from BRK (used for municipality boundaries)
- District- and neighborhood boundaries (municipalities are subdivided into districts, whereas districts are subdivided into neighborhoods

- Geografisch Basis Register (GBR)
- Kerncijfers Wijken en Buurten (table with demographic and socio-economic key figures of districts and neighbourhoods, for example percentage of male and female population, percentage of education level)

Bevolkingskernen – Urban agglomeration

The urban agglomeration product is based on the BBG, defining the contours of agglomeration areas according to certain preconditions without taking into account administrative borders. (www.cbs.nl)
Input datasets:

- BBG (Land use map)
- NWB (as a reference layer)

Grid-based statistics

The grid-based statistics have been introduced due to a need for standard contours that do not change over the years and provide better privacy protection. The standard grid sizes are 100 by 100 m and 500 by 500 m within RD New projection and 1 km * 1 km in ETRS89 (according to INSPIRE requirements). Statistical data are disseminated per grid, enabling better comparison per area. (CBS, 13). The following input datasets are used for the grid-based statistics:

- 100*100 m Grid feature classes
- 500*500 m Grid feature classes
- Kerncijfers Wijken en Buurten (table with demographic and socio-economic key figures of districts and neighbourhoods, for example percentage of male and female population, percentage of education level)

Figure 50- Grid-based statistics

Appendix 2: Description of selection of spatial datasets

In general, all used datasets are data covering the Netherlands and projected in RD New. Some datasets are very large and contain different tables (called "objects"). For the NWB, the road datasets are subdivided into Road, Water and Rail networks.

Name	Scale	Geometry type	Size (in MB or GB)	Data format	Objects
BRT	1:10.000	Varies per object	2,13 GB (FGDB), varies per year	FGDB or coverage	• Wegdeel • Spoorbaandeel • Waterdeel • Gebouw • Terrein • Inrichtingselement • Relief • Registratief gebied • Geografisch gebied • Functioneel gebied Additional: 59 relationshipclasses
BAG	1: 5000	Varies per object	10,4 GB (FGDB), 8,87 GB (BAG_01052011met relations.gdb	FGDB	7 objects (not all are used in one geoprocessing tool): • Panden • Nummeraanduidingen • Verblijfsobjecten • Staplaatsen • Ligplaatsen • Openbare Ruimte • Woonplaats Relationship classes: 8:
NWB	1:10.000	polylines	Not available	FGDB or coverage	• BWG Wegen (roads) • NWB vaarwegen (shipping routes) • NWB spoorwegen (rail roads)
BBG	1:20.000	Polygons	601 MB, varies per uear	FGDB or coverage	BBG contains one table, using 12 attributes
Wijk- en buurtkaart	1:	polygons	33,8 MB	FGDB or coverage	Contains one table, using 11 attributes

.

Appendix 3: Nomenclature logfiles

An explanantion of the nomenclature is provided in section 5.4 Setting up the test bed.

nomenclature by sandra desabandu

A = toolnamenr:
ArcGIS 9x; 10x vs ArcInfo Workstation 9.3 en 10.0 command
nr 0 = buffer
nr 1 = joinfield vs joinitem
nr 2 = dissolve
nr 3 = spatial join
nr 4 = intersect
nr 5 = split
nr 6 = clip
nr 7 = near
nr 8 = pointdistance
nr 9 = OD matrix

nr 10 = loop with cursor through records with Capatalize STREETNAME function
nr 11 = frequency
nr 12 = summarize statistics
nr 13 = union
nr 14 = Identity
nr 15 = Reselect - Feature class to feature class
nr 16 = Calculate
nr 17 = Add Join workaround

B = hardware:
0 = fat_client_normal
1 = big_data_inno
3 = laptop ArcGIS 10.2
4 = laptop ArcGIS pro

C = populationlength:
nr 0: municipality small
nr 1: municipality large
nr 2: province
nr 3: NL
nr 4: 10000 records
nr 5: 100000 records
nr 6: 500000 records
nr 6a: 600000 records
nr 6b: 700000 records
nr 6c: 800000 records
nr 6d: 900000 records
nr 7: 1000000 records
nr 8: 5000000 records
nr 9: 10000000 records
nr 10:50000000 records

D = populationwidth:
nr 0: only FID
nr 1: FID 2 attributes
nr 2: FID_all attributes
nr 3: synthetic (OID, shape)

E = index:
nr 0: not indexed
nr 1: indexed
nr 2: no spatial index
nr 3: spatial index default
nr 4: spatial index adaption

F = spatialsort:
nr 0: no sorting
nr 1: UL
nr 2: UR
nr 3: LR
nr 4: LL
nr 5: PEANO
nr 6: PEANO plus field

G = filetype:
nr 0: coverage info
nr 1: fgdb
nr 2: pgdb
nr 3: shp dbf

H = compressingcompacting
nr 0: not compressed not compacted
nr 1 : compressed
nr 2: compacted
nr 3: compressed and compacted

I = daytime:
nr 0: Monday office hrs
nr 1: Friday 7 pm

J = ArcGISversion:
nr 1: ArcGIS 10.1
nr 2: ArcGIS 10.0
nr 3: ArcGIS 10.2
nr 4: ArcGIS 9.3.1
nr 5
: Workstation 9
nr 6: Workstation 10
nr 7: ArcGIS Pro

#K = Processing:
nr 0: linear_default
nr 1: 64 bit background
nr 2: LAA
nr 3: parallel_2 threads
nr 4: parallel_4threads

#L = locationscript:

nr 0: network
nr 1: local

#M = coldwarmrun
nr 0 :cold
nr 1: warm

#N = RD new projection (since 03-10-2014)
nr 0: no projection
nr 1: RD new projection

#O = data
nr 0: real data
nr 1: synthetic data

#P = workspace
nr 0: different workspace
nr 1: same workspace

nomenclature
toolnamenr_ hardware_populationlength _populationwidth _index_ spatialsort_
filetype_compressingcompacting _daytime_ArcGISversion_Processing _ locationscript _ coldwarmrun

Appendix 4: Questionnaire geoprocessing and migration in Geo ICT Infrastructure

Contact data

Name of organization:

Name contact person:

Address:

Email:

Telephone number:

GIS users

1. Can you indicate how many of the GIS users are “light users” (map viewing, simple analysis, editing)
2. Can you indicate how many of the GIS users are heavy users (staff members that are tasked with heavy analysis of datasets, writing models) you have in your organisation?
3. Which GIS products are used :
    a. ESRI: Arc/Info Workstation/AML, ArcGIS Desktop/ArcPy, Arcview/Avenue
    b. Other: Open Source, Oracle, ETL

Geoprocessing & performance

1. Can you provide 1 or 2 examples of important GIS analysis products and the analysis steps that are taken?
2. Do you execute them via scripting, modelbuilder or on the fly?
3. What are the most important datasets you use for geoprocessing?
4. Which methods do you apply to improve performance or keep an acceptable level of performance? Examples: partitioning of datasets, parallel processing, etc.
5. Is there enough time available to experiment with performance improvement of geoprocessing?

ICT Infrastructure

1. How are the special needs of GIS users (need for higher calculation and storage capacity of the hardware, availability of up-to-date tooling) met by the ICT department of your organisation? For example, do you have staff members tasked with cooperation between ICT department and GIS users to ensure ?
2. How do you conduct the geoprocessing?
    a. On the local drive
    b. Via the network
    c. Virtual (if yes, which virtualization software, e.g. Citrix or VMWare)
    d. Other (e.g. cloud services)

Systems migration process (applicable for organizations that have migrated from an old GIS like Arc/Info Workstation to a more recent GIS)

1. From which system did you migrate?
2. What is the current system?
3. When did you migrate to the current system (e.g. ArcGIS Desktop)?

4. Did you redesign the geoprocessing scripts yourself or did you outsource it?
5. Did you observe a difference in performance and stability of the geoprocessing tools between old and new system? If yes, please answer following sub questions:
    a. Can you provide 3 examples?
    b. Did you anticipate on these differences before, e.g. via performance tests?
    c. If there is a negative change, explain how your organisation dealt with it? For example, did you have to stop the use of certain analysis models or did you have to re-develop them?
6. Do you still have GIS users that are working with the “old” system? If yes, please answer following sub questions:
    a. How many users?
    b. Why do they still use it?
    c. When and how will they migrate to the new system in the future?
7. What are the lessons learned from the migration project?
8. Are you planning to test new products (such as e.g. ArcGISPro)?

Appendix 5 Results interviews and questionnaires

GIS users

	Nr of light users (app.)	Nr of heavy users (app.)	GIS Software
Statistics Netherlands	**App. 25**	**App. 8**	**a. ESRI: ArcInfo Workstation/AML, ArcGIS Desktop/ArcPy, Arcview/Avenue, ArcGIS Server (for web services)** **b. Other: R Spatial (not used by Spatial Statistics, but methodology department), Quantum GIS (used by Spatial Statistics to view web services)**
Statistics Italy	50	20	a. ESRI: ArcGIS Desktop/ArcPy b. Other: Very limited use of Open Source
Statistics Portugal	25	5	*a.* ESRI Software: ArcGIS 10.2.2, ArcSDE 10.0 (Oracle 11g database), ArcGIS Server 10.0, ArcPAD 7.1, ArcView 3.3, ArcInfo workstation only rarely used b. *Other:* Oracle 11g database running ArcSDE 10.0, Quantum GIS, P-Mapper, Several python modules with geographical capabilities
Netherlands Environmental Assessment Agency	70	30	a. ESRI: ArcGIS desktop 10.1, ArcPy, ArcGIS Server (for web services) b. FME (trial), PostgreSQL/ PostGIS to import the BRK (Key Register Kadaster), experimental use of Manifold and QGIS
U.S. Geological Survey	20% (of the GIS users)[10]	80% (of the GIS users)	c. ESRI: ArcInfo Workstation/AML, ArcGIS Desktop/ArcPy, Arcview/Avenue d. Other: -

Table 25: GIS users of similar organizations

[10] The USGS has app. 10.000 staff members (scientists, technicians, support staff), figures of the total number of GIS users are not available

Geoprocessing and performance optimization

	Analysis examples[11]	Execution method	Datasets	Optimization methods	Time to experiment for optimization
Statistics Netherlands	• **BBG – Land use map for the Netherlands obtained from aerial imagery** • **Analysis of spatial features per municipality,** • **Municipality, - District- and Neighborhood based statistics** • **Grid-based statistics** • **Urban agglomeration areas**	• **Scripting** • **Modelbuilder**	• **National reference vector datasets (see chapter 2)** • **Imagery**	• **Partitioning** • **Local processing**	**No**
Statistics Italy	• Analysis of commuting flows • Automated Procedures for updating tabular data	• Scripting • Model builder	Network dataset	None	Not enough
Statistics Portugal	• Spatial analysis of building dataset • Scripts to validate and manage geographical data • Network analysis • Proximity statistics	• Applications • Scripting • Modelbuilder	• Geographical database with buildings (3.5 million records), • Street database • European GRID • other reference datasets	• Partitioning • Database versioning • Database schema for reference and dynamic datasets	Only during design process
Netherlands Environmental Assessment Agency	• Environmental models containing a large number of tools • Network Analysis)	• Scripting • Modelbuilder No performance difference has been detected between those methods	• National landuse data such as BBG, HGN, LGN • Imagery, AHN • Navteq data	• Multiprocessing for CPU bound tools (in ArcGIS) • Grid computing	Yes
U.S. Geological Survey	• Building elevation / hydro integrated data sets for hydro analysis • Raster analysis	On the fly	Land use and land cover, elevation	Mostly Tiling of data	Yes

Table 26: Geoprocessing and perfomance optimization at similar organizations

[11] For SN, see also section 2.2

ICT Infrastructure

	Communication about ICT needs of the GIS users	Which part of the infrastructure is used for geoprocessing
Statistics Netherlands	• **1 (parttime) ICT coordinators for the GIS staff** • **2 application managers of the ICT department with knowledge of GIS**	• **Fat client** • **Network (script and user profile)**
Statistics Italy	• GIS data is stored locally and often administered by a GIS manager. Since data is administered, edited, and used only by a few people who are closely working together, problems such as concurrent data use conflicts can be resolved easily. • Data distribution is also simple because the needs of the user community are known and the GIS group does not need to deal with a wide variety of data requirements.	• Data are stored in a local or shared drive as file-based datasets such as shapefiles or in a personal geodatabase
Statistics Portugal	• Communication of special needs to the ICT department e.g. regarding disk-space and processing capabilities • Senior staff of the Geo-Information Unit are expected to cooperate closely with the ICT department	Recurrent processes are executed with data on our network, or on our local drive and the result is copied to our network or our geographical database.
Netherlands Environmental Assessment Agency	• Heavy GIS users are represented in a GIS expert group • There is a special department called “Information, data and methods” (IDM) which supports other departments concerning information management, data and methods. The IDM Geo point has been established for GIS users	• Storage of data in fgdb format on the network, dissemination of the data via layerfiles on a data portal for the users. • The PBL works with virtual Citrix desktops, however, users can choose whether they want to use the local desktop or the virtual desktop • The data centre is situated on the terrain of the main location, containing the virtual servers but also the “physical” ArcGIS servers. The ArcGIS servers can be used as a computer grid
U.S. Geological Survey	Generally these needs are met as they are critical to the research of USGS	All options are used (local processing, network, virtual desktop and cloud processing)

Table 27: ICT infrastructure similar organizations

Migration

organization	Former system	Current system	Periode of migration to Desktop	Outsourcing	Remaining workstation users	Motivation to work with Workstation	Migration last users	Difference WS and desktop
Italy	ArcInfo Workstation	ArcGIS Desktop 10.1	2002	ESRI performed a number of initial tasks such as the (file geo) database design and a pilot project	A few users tasked with editing	No information	No information	• ArcInfo Workstation difficult to maintain • No possibility to document meta data in ArcInfo Workstation
Portugal	ArcInfo workstation and ArcView	ArcGIS desktop, ArcSDE and ArcGIS Server	2003-2007	Only help from ESRI to create first version of geographical database in 2003 and to develop the applications for the 2011 census	A few users of ArcView 3.3.	Produce thematic maps, feel more comfortable	Possibility of making thematic maps will continue, only editing within geodatabase not recommended	• The new geoprocessing tools are more powerful and flexible than the old AML scripts. • It is easier to link geographical to alphanumeric data. • No complaints about the performance with ArcSDE.[12].
Netherlands Environmental Assessment Agency	ArcInfo workstation	ArcGIS Desktop 10.1	No information	Yes, the AML scripts of the models have been redesigned by a company to Python.	No remaining user	Not applicable	Not applicable	All models have been tested on stability – during that process a number of models have been discarded because they were not running stable in Python/ArcPy.
U.S. Geological Survey	ArcInfo Workstation	ArcGIS Desktop 10.1, sp1	2001	No	Handful	The application still works satisfactory (“still works “)	Gradual process, using both applications	Some things were faster some were not as good. 10.1 provided same (and improved_ functionality with arcpy.mapping, model builder and python capability

Table 28: Migration experiences form similar organizations

[1212] Datasets of about 3,5 million records can be used without any problems (although execution time not known)

Appendix 6: Error Report ESRI

Van: Esri Nederland Support [mailto:support@Esri.nl]
Verzonden: donderdag 9 april 2015 10:53
Aan: Zuurmond, Drs. M.
Onderwerp: 1057606 - Verschillend resultaat bij zelfde tool, machine, data, etc.

Geachte heer Zuurmond,

Bedankt voor uw supportaanvraag bij Esri Nederland. Uw vraag met de omschrijving: *"Verschillend resultaat bij zelfde tool, machine, data, etc."* is geregistreerd in ons systeem onder nummer **1057606**. Een van onze support engineers zal zo spoedig mogelijk contact opnemen met een antwoord of aanvullende vragen.

Klik hier voor details over uw supportvraag.

Heeft u nog geen inloggegevens voor mijn.Esri.nl, dan kunt u deze aanvragen via http://mijn.Esri.nl/standaard/wachtwoord-aanvragen. Via het portaal Mijn Esri kunt u zich - naast supportvragen insturen - inschrijven voor events. Ook kunt u uw mailingbehoefte aanpassen naar uw voorkeuren. Deze pagina bereikt u direct via https://mijn.Esri.nl/aanpassen-persoonsgegevens-prs.

Met vriendelijke groet,

Esri Nederland Support

T: +31 (0)10 217 07 50
E: support@Esri.nl

Originele vraag / Original question:

Beste mensen van support,

Bij ons heeft afgelopen jaar een afstudeer onderzoek van stagiaire Sandra Desabandu naar performance van GIS software plaats gevonden.
Een aspect dat tot onverwacht resultaat dat naar voren kwam was dat bij twee tools die elk 5 keer in 1 loop zijn gerund verschillende output gegenereerd werd.
•Bij een dissolve tool op een selectie van BBG (alle werd in de 1 tm 3 en 5 loog een output gegenereerd 150074 records en in de 4e loop een output van 150087 records.
•Bij een intersect tussen Bos vlakken (bbg = 60) en nwb werden 1e, 2e en 5e run 85488 records (intersect lijnstukken) gegenereerd; in de 3e en 4e run 85527 records.

Gebruikt zijn AGDesktop 10.1 sp1, Windows 7 sp1

Hoewel de verschillen minimaal zijn, is het toch raar dat met dezelfde tool in een python script met loop verschillende output wordt gecreëerd. Output is elke keer met een volgnummer weggeschreven.
Is het bekend bij ESRI dat bij deze twee tools dit op kan treden? Eventueel kan ik een dvd sturen met data en script.

Mvg
Marijn Zuurmond

REFERENCES

Abdelguerfi, Mahdi; Mahadevan, Venkata; Challier, Nicolas; Flanagin, Maik; Shaw, Kevin & Ratcliff, Jay. (2005). A High Performance System for Processing Queries on Distributed Geospatial Data Sets. In M. Daydé, J. Dongarra, V. Hernández & J. L. M. Palma (Eds.), *High Performance Computing for Computational Science - VECPAR 2004* (Vol. 3402, pp. 119-128): Springer Berlin Heidelberg.

Aji, Ablimit; Wang, Fusheng; Vo, Hoang; Lee, Rubao; Liu, Qiaoling; Zhang, Xiaodong, & Saltz, Joel. (2013). Hadoop GIS: a high performance spatial data warehousing system over mapreduce. *Proc. VLDB Endow., 6*(11), 1009-1020. doi: 10.14778/2536222.2536227

Al-Azzoni, Issam; Zhang, Lei, & Down, Douglas G. (2011). *Performance evaluation for software migration*. Paper presented at the Proceedings of the 2nd ACM/SPEC International Conference on Performance engineering, Karlsruhe, Germany.

Alsultanny, Yas. (2010). Database management and partitioning to improve database processing performance. *Journal of Database Marketing & Customer Strategy Management, 17*(3-4), 271-276. doi: 10.1057/dbm.2010.14

Practical Performance Analyst. (2014a). Performance Engineering Fundamentals | Performance Testing 101. from http://www1.practicalperformanceanalyst.com/fundamentals-of-performance-capacity-management/performance-engineering-fundamentals-performance-testing-101/#sthash.kDdVjBWE.dpbs

Practical Performance Analyst. (2014b). Performance Engineering Fundamentals | Workload Modeling 101. 2014, from http://www1.practicalperformanceanalyst.com/fundamentals-of-performance-capacity-management/performance-engineering-fundamentals-workload-modeling-101/

ASPRS. (2012). LASer (LAS) File Format Exchange Activities. 2015, from http://www.asprs.org/Committee-General/LASer-LAS-File-Format-Exchange-Activities.html

Batcheller, J.K.; Gittings, B.M.; Dowers, S. (2007). The performance of vector oriented data storage strategies in Esri ArcGIS. *Transactions in GIS, 11*(1), 47–65.

BBC. (2015). CPU and memory. *Bitesize.* 2015, from http://www.bbc.co.uk/education/guides/zmb9mp3/revision/2

Bell, John T. Dr. (2013). Virtual Memory. 2015, from http://www.cs.uic.edu/~jbell/CourseNotes/OperatingSystems/9_VirtualMemory.html

Bogdan Simion, Suprio Ray, Angela Demke Brown (2012). *Surveying the Landscape: An In-Depth Analysis of Spatial Database Workloads*. Paper presented at the 20th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL GIS 2012), Redondo Beach, California, US.

Bouckaert, Stefan ; Vanhie-Van Gerwen, Jono; Moerman, Ingrid; Phillips, Stephen C ; Wilander, Jerker; Rehman, Shafqat Ur; Turletti, Thierry. (2011). Benchmarking computers and computer networks.

Chen, Hue-Ling, & Chang, Ye-In. (2005). Neighbor-finding based on space-filling curves. *Inf. Syst., 30*(3), 205-226. doi: 10.1016/j.is.2003.12.002

Childs, Colin. (2009). The top nine reasons to use a file geodatabase. *ArcUser*.

Computerhope. (2015). What is the difference between a 32-bit and 64-bit CPU? , 2015, from http://www.computerhope.com/issues/ch001498.htm

Desabandu, Sandra, & Eijkelenboom, Ernst (2014, 05.11.2014). [Verslag meeting ESRI dd 5 november 2014].

Desabandu, Sandra, & Emons, Maarten. (2014). Interview with Maarten Emons dd 15 december 2014.

Desabandu, Sandra; Put, Arjan van der, & Spoon, Martijn (2014). [Verslag gesprek bij het Planbureau voor de leefomgeving dd 2 mei 2014].

Desabandu, Sandra; Keuren, Anke, & Scheper, Eddy. (2014). Interview with Aris dd .4 juni 2014.

Elkins Jr, Rob, & Macleod, Charlie. (2014). *ArcGIS Pro .NET SDK: The Road Ahead.* Paper presented at the ESRI US 2014.

Elsevier. (2012). Gids de beste gemeenten : Leeswijzer bij ranglijst gemeenten en buurten. *Elsevier,* 80, 81.

ESRI.(2013). About compressing file geodatabase data. 2014, from http://webhelp.esri.com/arcgisserver/9.3/java/index.htm#geodatabases/about_c1145393054.htm

Esri. (2004). Overview of ArcGIS Topology. *ESRI White Paper*.

ESRI. (2010a). Computations on vector data using a GPU. 2015, from http://blogs.esri.com/esri/apl/2010/03/30/computations-on-vector-data-using-a-gpu/

ESRI. (2010b). Coverage data limitations. 2014, from http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#//00140000001r000000.htm

ESRI. (2010c). Coverage topology. 2014, from http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#//001400000016000000.htm
ESRI. (2012). Troubleshooting Slow Performance in ArcGIS for Desktop. 2014, from http://blogs.esri.com/esri/supportcenter/2012/06/07/troubleshooting-slow-performance-in-arcgis-desktop/
ESRI. (2013a). Compacting file and personal geodatabases. 2014, from http://resources.arcgis.com/en/help/main/10.1/index.html#/Compacting_file_and_personal_geodatabases/018s0000000s000000/
ESRI. (2013b). Compressing file geodatabase data. 2014, from http://resources.arcgis.com/en/help/main/10.1/index.html#/Compressing_file_geodatabase_data/018s00000003000000/
ESRI. (2013c). Dissolve (Data Management). from http://resources.arcgis.com/en/help/main/10.1/index.html#/Dissolve/00170000005n000000/
ESRI. (2013d). Frequency (Analysis) 2014, from http://resources.arcgis.com/en/help/main/10.1/index.html#/Frequency/00080000001w000000/
ESRI. (2013e). How Intersect works. 2014, from http://resources.arcgis.com/en/help/main/10.1/index.html#//00080000000z000000
ESRI. (2013f). Near (Analysis) 2014, from http://resources.arcgis.com/en/help/main/10.1/index.html#//00080000001q000000
ESRI. (2013g). A quick tour of setting a spatial index. 2014, from http://resources.arcgis.com/en/help/main/10.1/index.html#//003n0000001r000000
ESRI. (2013h). Summary Statistics (Analysis). 2014, from http://resources.arcgis.com/en/help/main/10.1/index.html#/Summary_Statistics/00080000001z000000/
ESRI. (2013i). Union (Analysis) 2014, from http://resources.arcgis.com/en/help/main/10.1/index.html#/Union/00080000000s000000/
ESRI. (2014a). ArcInfo Workstation Product Life Cycle Status.
ESRI. (2014b). Parallel Processing factor (Environment setting). 2014, from http://resources.arcgis.com/en/help/main/10.1/index.html#//001w0000004m000000
ESRI. (2014c). The spatial grid index. 2014, from http://resources.arcgis.com/en/help/main/10.1/index.html#//006z0000002n000000
ESRI. (2014d). What's new in ArcGIS 10.2.1. Retrieved 14.01.2015, 2015, from http://resources.arcgis.com/en/help/main/10.2/index.html#//016w0000005v000000
ESRI. (2014e). What is lidar data? , 2015, from http://resources.arcgis.com/en/help/main/10.2/index.html#//015w00000041000000
ESRI. (2015). Projects in ArcGIS Pro. Retrieved 05.05.2015, 2015, from http://pro.arcgis.com/en/pro-app/help/projects/overview/what-is-a-project.htm
ESRI. (unknown). Infrastructure Performance Considerations. 2014, from http://resources.arcgis.com/en/communities/enterprise-gis/01n200000021000000.htm
ESRI/VMWare. (2013). Esri® ArcGIS® 10.1 for Server, Esri ArcGIS 10.2 for Server on VMware® vSphere.
Frank Pizzi, Andrew Sakowicz. (2013). *Technical workshop. ArcGIS Enterprise Systems: Performance and Scalability.* Paper presented at the Esri International User Conference, San Diego, California.
Godfrind, Albert. (2008). Introduction to performance of Oracle Spatial databases: Oracle gebruikersclub Holland.
Goedhuys, Mariette. (2014). *HGA fase 8 – Vooronderzoek ArcView 3.1 en PX-Map (concept)*. Statitics Netherlands.
Guan, Weihe. (2006). ESRI Products: Center for Geographic Analysis Harvard University.
Harley, Mark, & Gellerstedt, Brad. (unknown).The Role of Grid Size Optimization in ArcSDE Performance Tuning.
Hartling, Ken. (2012). Be successful overlaying large, complex datasets in Geoprocessing. 2014, from http://blogs.esri.com/esri/arcgis/2012/06/15/be-successful-overlaying-large-complex-datasets-in-geoprocessing/
Hennessy, John L. , & Patterson, David A. (2007). *Computer architecture : a quantitative approach* (4 ed.). San Francisco: Morgan Kaufmann Publishers.

Hoel, Erik G., & Park, Mike. (2013). *Big Data: Using ArcGIS with Apache Hadoop.* Paper presented at the ESRI International Developer Summit, Palm Springs, CA.
Huber, William A. (2011). Does using RAM Disk improve ArcGIS Desktop performance appreciably? *GIS Stackexchange.* 2014, from http://gis.stackexchange.com/questions/537/does-using-ram-disk-improve-arcgis-desktop-performance-appreciably
Iparraguirre, Edgar Walter. (2014). ESRI health check CBS.
Janakiram, MSV. (2012). What is Common between Mumbai Dabbawalas and Apache Hadoop? , 2015
Kaler, Tim. (2012). Spatial Data Structures - Performance Comparision. In O. Moll (Ed.).
Klinkenberg, Brian. (1997). Unit 63 – Benchmarking. Retrieved 01.10.2014, 2014, from http://ibis.geog.ubc.ca/courses/klink/gis.notes/ncgia/u63.html
Kouyoumjian, Victoria. (2010). The New Age of Cloud Computing and GIS. *ArcWatch*. http://www.esri.com/news/arcwatch/0110/feature.html
Lee, Dan, & Hardy, Paul. (2006). *Design and Experience of Generalization Tools.* Paper presented at the AutoCarto Vancouver.
Manifold.(2015). Homepage Manifold. 2015, from http://www.manifold.net/
Matty, Shamal Kiran. (2012). *Comparative Study of Oracle Spatial and Postgres Spatial.* San Diego State University.
Menasce, Daniel A., & Almeida, Virgilio. (2001). *Capacity Planning for Web Services: metrics, models, and methods*: Prentice Hall PTR.
Morais, C. Dempsey. (2008, 20.07.2014). Why ArcView 3.x is still in use. *GIS Lounge.* Retrieved 20.07.2014
Morehouse, Scott. (1989). *The architecture of Arc/Info.* Paper presented at the Proceedings of the Auto Carto 9 Conference, Baltimore, MD, American Society for Photogrammetry and Remote Sensing/American Congress for Surveying and Mapping, Baltimore.
Mullins, Craig S. (2010). Defining Database Performance. 2014, from http://www.dbta.com/Columns/DBA-Corner/Defining-Database-Performance-70236.aspx
Statistics Netherlands. (2015a). About us. 2015, from http://www.cbs.nl/en-GB/menu/organizatie/default.htm?Languageswitch=on
Statistics Netherlands. (2015b). Bescherming persoonsgegevens. 2015, from http://www.cbs.nl/nl-NL/menu/organizatie/bescherming-persoonsgegevens/default.htm
Oosterom, Peter van. (1999). Spatial Access Methods. In G. Longley, Maguire en Rhind (Ed.), *Geographical Information Systems Principles, Technical Issues, Management Issues, and Applications* (pp. 385-400): Wiley.
Opperdoes, Eddy (2014). [Communication with Eddy Opperdoes via email in 2014].
Pardy, Jason, & Hartling, Ken. (2013). *Efficient data management and analysis with geoprocessing.* Paper presented at the Esri International User Conference.
Paton, NormanW; Williams, M. Howard; Dietrich, Kosmas; Liew, Olive; Dinn, Andrew, & Patrick, Alan. (2000). VESPA: A Benchmark for Vector Spatial Databases. In B. Lings & K. Jeffery (Eds.), *Advances in Databases* (Vol. 1832, pp. 81-101): Springer Berlin Heidelberg.
Paul, Subharthi (2008). Database systems performance evaluation techniques. Retrieved 09.12.2013, 2013, from http://www.cs.wustl.edu/~jain/cse567-08/ftp/db/
Performance Team , Windows Server (2009). Interpreting CPU utilization for performance analysis. 2014, from http://blogs.technet.com/b/winserverperformance/archive/2009/08/06/interpreting-cpu-utilization-for-performance-analysis.aspx
Peters, Dave. (2008). *Building a GIS: System Architecture Design Strategies for Managers* (1 ed.): ESRI Press.
Peters, Dave. (2014). The Evolution of GIS Software. 2015, from http://blogs.esri.com/esri/esri-insider/2014/12/22/the-evolution-of-gis-software/
Qian, Liujian and Peuquet, Donna J. (1997). *A Systematic Strategy for High Performance GIS.* Paper presented at the Annual Convention and Exposition Technical Papers Seattle, Washington
Ray, S., Simion, B., & Brown, A. D. (2011, 11-16 April 2011). *Jackpine: A benchmark to evaluate spatial database performance.* Paper presented at the Data Engineering (ICDE), 2011 IEEE 27th International Conference on.
Saalfeld, Alan. (1998). Organizing spatial data for spatial analysis.
Sakowicz, Andrew, & Pizzi, Frank. (2014). *ArcGIS for Server performance and scalability: Testing methodologies.* Paper presented at the Esri International Developer Summit.
Schets, Ian, & Desabandu, Sandra (2014). [Communication with Ian Schets].
Simion, Bogdan, Ray, Suprio, & Brown, Angela Demke. (2012). Speeding up Spatial Database Query Execution using GPUs. *Procedia Computer Science, 9*, 1870-1879. doi: 10.1016/j.procs.2012.04.205

Sorokine, A; Myers, A; Liu, C; Coleman, P; Bright, E; Rose, A' &. Bhaduri, B. (2012). *Tackling BigData: Strategies for Parallelizing and Porting Geoprocessing Algorithms to High-Performance Computational Environments*. Paper presented at the GIScience 2012

Stackexchange. (20111-2014). ArcGIS geoprocessor memory leak: what's the underlying cause and why is it so hard to fix? , 2014, from http://gis.stackexchange.com/questions/20225/arcgis-geoprocessor-memory-leak-whats-the-underlying-cause-and-why-is-it-so-ha

Stackoverflow. (2009). CPU bound and I/O bound? , 2014

Stackoverflow. (2012). Which is the relationship between CPU time measured by Python profiler and, real, user and sys time? Retrieved 01.10.2014, 2014, from http://stackoverflow.com/questions/9533179/which-is-the-relationship-between-cpu-time-measured-by-python-profiler-and-real

Stormen, T.M. (2013). CBS Infra-overzichtsplaat.

Teachbook. (2012). The memory hierarchy. 2014, from http://blog.teachbook.com.au/index.php/2012/02/memory-hierarchy/

Theobald, David M. (2001). Topology revisited: representing spatial relations. *International Journal of Geographical Information Science, 15*(8), 689-705. doi: 10.1080/13658810110074519

Tijssen, Theo , Quak, Wilko , & van Oosterom, Peter. (2012). Geo DBMS als standaard bouwsteen voor Rijkswaterstaat. Delft: OTB Research Institute for Housing, Urban and Mobility Studies, TU Delft.

Princeton University. (2005). Lecture 20: Input/Output: Princeton University.

Van Dyke, Heather Ann (2009). [ESRI Coverage Model].

van Oosterom, Peter; Martinez-Rubi, Oscar; Ivanova, Milena; Horhammer, Mike; Geringer, Daniel; Ravada, Siva, & Gonçalves, Romulo. (2015). Massive point cloud data management: Design, implementation and execution of a point cloud benchmark. *Computers & Graphics*(0). doi: http://dx.doi.org/10.1016/j.cag.2015.01.007

Wikipedia. (2015). Computer Performance. from http://en.wikipedia.org/wiki/Computer_performance

Wilson, Tom. (2010). Principles of performance measurement. *CMG MeasureIT*.

Woodside, M., Franks, G., & Petriu, D. C. (2007, 23-25 May 2007). *The Future of Software Performance Engineering.* Paper presented at the Future of Software Engineering, 2007. FOSE '07.

Zhang, Jianting, & You, Simin. (2012). *CudaGIS: report on the design and realization of a massive data parallel GIS on GPUs*. Paper presented at the Proceedings of the Third ACM SIGSPATIAL International Workshop on GeoStreaming, Redondo Beach, California.

Zhao, Gang; Bryan, Brett A.; King, Darran; Song, Xiaodong, & Yu, Qiang. (2012). Parallelization and optimization of spatial analysis for large scale environmental model data assembly. *Computers and Electronics in Agriculture, 89*, 94-99. doi: 10.1016/j.compag.2012.08.007

Zuurmond, Marijn. (2013). Print Screen of a script. Statistics Netherlands