Centraal Bureau
voor de Statistiek

Vision

Methodology Research Vision 2020-2025

F.P. Pijpers

Table of contents

1. Introduction

Statistics Netherlands wishes to continue to play a leading role in the publication of trustworthy and comprehensive statistical information, responding to the needs of society, and also continue to guarantee the high quality of this statistical information. Society changes in an ever increasing tempo, and therefore Statistics Netherlands needs to be able to adjust to these changes to remain relevant, which requires innovation. Statistics Netherlands has laid down its ambitions for the coming five years in recent whitepapers, as well as laying down targets for improvement of its processes in a rolling strategic agenda. These targets have been further specified and expanded upon in three prioritary innovation programmes and a data strategy, which have been consulted and considered carefully in the process of compiling this research vision.

The various statistical divisions provide society with the statistical information it needs. In order to do this properly they need to have access to the correct methodology. The work of the division Dataservices, Research and Innovation (DRI) is organised to respond to the needs of the statistical divisions as well as ensuring that the full methodological instrumentarium remains up-to-date. As the need arises DRI itself also takes the initiative for carrying out research which is part of exploring and expanding the horizons of official statistics.

The research programme methodology has the aim for Statistics Netherlands to be recognised internationally as a gold standard for the manner in which official statistics are compiled: Testable, Reproducible, Ethical, and Efficient. The operationalisation of these four desired characteristics of methods and ways of working requires research and development in a number of areas: the expertise themes described in this plan.

1.1. Trends in society

Data based society

Over the most recent decennia our environment, within the home and the workplace as well as more widely in the public space, has become pervaded with sensors and other types of modern technology which collect data. Sometimes we are quite conscious of data collection, such as whenever we share information about ourselves or our friends and acquaintances via social media. However, even more data is generated by a variety of equipment and apparatuses, intended in the first place to regulate and control the correct functioning and maintenance of that equipment itself, or for instance to provide feedback to manufacturers. The infrastructure which collects such data and stores these centrally is often referred to as the 'internet of things' (Ashton, 2009). Regardless of the original purpose for the registration and archiving of data, any data can be used for other purposes than originally intended. For private parties such as large technology and media companies, data are a potential source of income; hence the expression "data are the new oil" (ascribed to Clive Humby in 2006).

There is a downside in the sense that repressive governments or organisations can (ab)use data and technology in order to suppress opinions or behaviours that it considers subversive or otherwise. The same technology which can assist every world citizen to retrieve information simply and quickly can also be misused by radical or violent groups to disseminate propaganda or

misinformation. From this it is evident that the 'big data' which is so freely available via the internet, does not automatically or self-evidently have genuinely useful veracious information content.

There is a trend which is gathering pace in Dutch government towards data driven decision-making. Intelligent analysis of data can give rise to improvements in the formulation of policy or particular measures, which also allow for a proper evaluation of such policies at a later stage. In addition, data can help to enable implementing or operating policies more effectively. The development of data driven policymaking is visible at the level of central government: e.g. regarding whether and what subsidies that are provided are in fact effective and which are not. However also more local, lower level, governments are implementing similar approaches to policy making: here one can think of e.g. city deals where town councils use data to enable targeting their resources primarily to their most vulnerable or in need residents, or in order to get a clear view of 'undermining' activities where (organised) criminality attempts to subvert or influence legal activities including (local) government.

For institutions that have been created for the public good, such as Statistics Netherlands, freely available data can be a source of information that is more responsive or faster in registering trends or changes in society. To utilize this, it is necessary that Statistics Netherlands is capable of distinguishing data that are and that are not useful for this purpose. Correct information needs to be extracted from the gigantic flow of (un-)structured data produced by people and technology: this includes for instance photographs and texts alongside more traditional numeric data. Perhaps of even greater importance for Statistics Netherlands is that there needs to exist an ethical foundation for the data collection and processing (see e.g. Floridi & Taddeo, 2016): a common beneficial societal goal needs to be served by the collection, analysis and publication of the information. Every possible precaution must be taken to ensure that the analysis process is not prejudiced or biased. Transparency of any process is an essential aspect of this. The publications of Statistics Netherlands must remain accessible to as wide as possible an audience of all citizens of the Netherlands, while at the same time taking care that their privacy is preserved

Closing the information divide: from data to statistics

The developments of broader society to become ever more data-dominated, poses a number of challenges for Statistics Netherlands. People and companies make ever greater amounts of information about themselves available all over the world via the internet, with very few restrictions. Despite this, the willingness to participate in the surveys that Statistics Netherlands traditionally organises as part of its primary data collection continues to decline. Statistics Netherlands has played a trail blazing role in combining data from (public) registers, from unstructured 'web-scraping' data, and from surveys having been taken either in person, by telephone, or via the web. It remains a priority to be able to correct for selectivity and bias specific to each of these modes or channels through which the data are obtained. The aim is to further advance the reduction of the burden of surveys wherever and whenever possible.

Increasingly, there is a call that the data generated by people and companies must be regarded as a type of property which has a certain value. In line with this thinking is that Statistics Netherlands, in its relationship with companies, has a quid pro quo arrangement that in return for data being provided by the companies they get access to relevant statistical information, for instance concerning the business sector(s) in which they are active. For its contact with the general public it is important that Statistics Netherlands can be more responsive, i.e. faster and more interactive, when queried. A very good example of this is to have implemented an automated virtual assistant: an 'information dialogue' which supports accessing the large library of 'open data'. It is clear that such an information dialogue also creates new challenges for Statistics Netherlands. It is necessary for Statistics Netherlands to have the ability 'on the fly' to prevent unintended breaches of privacy or other disclosure of sensitive data, by means of researching and implementing appropriate techniques.

A different way to improve and facilitate collecting data is to lower the threshold for participation by means of a degree of 'integration by design'. An app, downloaded and installed on a smartphone, which helps people to manage their own household budget can also, if and when consent is given by the user, be utilised to provide Statistics Netherlands with appropriate and relevant summary data for its ongoing household budget statistics. Ideally these data would be streamed directly into the production pipeline for these statistics. Once the process of the collection of data, as well as the data cleaning and analysis process, are set up well this implies an increase in efficiency as well as accuracy. Even in circumstances where the data collection itself is not managed by Statistics Netherlands, but an external provider has data available, it is possible to create a data-interface. Through the interface only that sector of the data which is necessary for the statistical process is brought into Statistics Netherlands, perhaps even already aggregated to some extent. Such filtering of a data stream before further processing and analysis is a form of 'edge computing' which has clear benefits in terms of efficiency improvement.

For local and national governance as well, it is necessary to have access to the right data and the knowledge to interpret such data correctly. In practice in particular local governments run into difficulties with this. One of these is for example the issue of proportionality and purpose in the collection of data. Regulations such as the GDPR directives place clear constraints which have the paramount purpose of protecting privacy of citizens. As a consequence it is necessary to develop new techniques to be able to share information whilst ensuring that such protections remain in place. Transparency requires that those data that can be provided in the form of open data are provided, and in a form that optimally supports the needs of society. Here too Statistics Netherlands sees a role for itself in developing appropriate methods and implementing techniques.

Another way to enhance the spectrum of applications of statistics is model-based estimation, for instance in order to produce *real time statistics*, as has already been noted in a report by Bakker (2013).
Here, real time statistics are to be interpreted in the sense that the (usually) monthly cadence with which data is collected also leads to monthly publication of updates to official statistics with a delay of preferably no more than one

month compared to the date at which the data collection was completed. While a cadence of more than once per month is not unthinkable for some statistics, at the moment there does not appear to be an evident need or demand to increase publications to such a fast rate. The data for many statistics are in principle available *real time* in the cases where Statistics Netherlands itself is in charge of the primary data collection. Normally however, the completeness and quality of data leave something to be desired until some time into the data collection process. There is a significant public interest in timely publication of important indicators such as economic growth rates, consumer confidence, unemployment, or well-being. Several of the core target audiences for these official statistics, such as governance and industry, are being served by these indicators. Problematic is that a number of these fast first estimates or 'flash estimates' have a tendency to deviate from the more robust, but slower versions of these indicators, published at a rather later stage. To some extent such discrepancies can be predicted by appropriate methods, and therefore can be solved during the processing of the data, which requires some investments. Where the conflicts between 'flash' and 'definitive' estimators cannot be solved in this way, more attention needs to be paid to the communication of the tightrope balance in the two types of measurements that needs to be maintained, between speed and accuracy.

From statistics to meaning

Were Statistics Netherlands to limit itself to the publication of official statistics in the form of tables, it would insufficiently fulfil its public role in the way that policy makers and the general public have a right to expect. In journalism there is already a widespread service of 'fact checking' columns and features. In these the pronouncements of policymakers, politicians, and other notable people or institutions with a public role are being checked for their correctness. In such features the relevant statements are being placed into context as well and a certain amount of clarification is provided in the 'translation' to quantitative numeric from. Statistics Netherlands has shown through its novel publication strategy in the past few years that it is also well placed to provide a broader context to the bare quantitative data. In this way a clear technical interpretation and meaning converts the bare data into information, without taking a particular political or societal stance.

In an era in which 'big data' is widely and easily available, there is an increasing risk that spurious relationships or correlations appear or seem to be confirmed. For private parties this might well be sufficient, if their primary purpose is to maximise an annual profit for themselves or their customers. For the purposes of official statistics a similar nonchalant attitude towards causal relationships is unacceptable. One important reason for this is that Statistics Netherlands has the responsibility to maintain consistent time series and statistics over time scales of many decades. Another reason is that using or reporting relationships without sufficient foundation, and which in the longer term turn out to be false, may well have ethically unacceptable consequences.

Quite often Statistics Netherlands limits itself to the generically true statement that correlation does not imply causation. Often, no more is possible in the

working environment of official statistics. Almost by definition in this field it is impossible to conform to the demands and constraints of controlled scientific experiments set up to identify causal relationships. However, in some cases it is possible to perform statistical tests of statements or hypotheses concerning putative causality which allow falsification. In the sense proposed by Popper (1934, 1956) this is in fact the only correct way of proceeding which is in line with the scientific method. It is an important component of the societal mission of Statistics Netherlands, to apply our knowledge and expertise to provide *whenever possible* such technical interpretations and falsification studies. In order to be able to do this it is necessary to follow closely the scientific developments in the area of techniques used to falsify hypotheses of causal relationships in a complex society. In this way, such tools can and must become part of the arsenal of techniques that Statistics Netherlands has at its disposal.

International context

The wide range of topics about which Statistics Netherlands publishes data must be seen in the context of the position of the Netherlands as member of the EU and as international trade and transport hub. Over time there is a transfer of tasks, responsibilities, and powers from Dutch governance to the EU. Also the trend continues towards further global internationalisation of trade and industry. Not only is the number and the influence of multinational companies increasing, more and more goods are traded via Internet, for which borders barely exist. This means that for some official statistics it makes sense to compile them for the EU as a whole, or at the very least in a way that is harmonised across the EU. A part of such harmonising efforts is the concept of *open data* and exploring the possibilities for developing a common infrastructure for European public data and official statistics (Hulliger *et al.*, 2012). This context reinforces the need for improved methods in the area of data-integration but also in the area of data-security.

1.2. From societal trends to themes of research

Statistics Netherlands' has the mission to provide a good service, tailor-made for local, regional, as well as national government. This implies that it is capable, technically and methodologically, of presenting facts in a manner which corresponds closely to the information needs of stakeholders. Objectivity in information provision does not equate to exclusively presenting tables of aggregated data. For a factual foundation of policies and decision-making it is important that Statistics Netherlands brings to the foreground the information which may not be immediately apparent in the numbers. Differences between various groups in the population, trends in time, or breaks in those trends, regularly are interpreted outside of Statistics Netherlands as if they are the consequence of certain policies or regulations, or the absence thereof. With the expertise and data which Statistics Netherlands has at its disposal, some of such putative *causations* can be excluded. This is a component of the research theme complexity science (see section 5.6). It is important that Statistics Netherlands

not only publishes the numbers and the trends, but is also proactive in explaining the meaning or implications of such data by well-founded technical analyses. The statistical techniques (sections 5.5 and 5.6) with which such analyses are performed are core results from the methodological research-programme.

The first paragraphs of section 1.1 mention the issues around separating appropriate from inappropriate data and information-extraction, in particular for the research of Statistics Netherlands within the theme Big Data, Data Mining and Artificial Intelligence (section 5.2). In some cases it will be important to be able to already filter and select data out of vast external data flows, even before any data is ingested in the databases maintained by Statistics Netherlands. As a consequence improvements are necessary in data processing (section 5.7). There is a well-known reduction in the willingness to participate in surveys, already mentioned in the paragraphs on closing the information divide. As a consequence it is necessary to investigate new observing techniques (section 5.1), and also to investigate integration of data from all the diverse channels and sources (section 5.3). In order to improve, in return, the provision of data to all stakeholders and society at requires research in the area of data querying (section 5.7). In returning data, the privacy of citizens must continue to be accounted for properly, and parallel to this it is necessary to prevent inadvertent publication of sensitive information concerning the daily order of business of individual companies. This requires continued research in the area of data security (section 5.4).
Experience from the methodological research of the past decennia has taught that for virtually every research project being carried out, expertise is necessary from more than just one of the research themes. In practice there is an overlap in the application of the themes, and hence there is a close collaboration between methodologists as well as between the methodology teams on the one hand and other divisions of Statistics Netherlands on the other hand.

The strategic agenda of Statistics Netherlands features innovation very high on its list of priorities. Methodological development plays a key role in innovation. The vision for methodological research presented in this document provides directions for methodological improvements as well as completely new methods and techniques. These are necessary to provide optimal support for the strategic aims of Statistics Netherlands. In the detailed descriptions of the themes in the research programme, wherever relevant an explicit mention is made of the specific strategic aims that are being supported.
The process flow within Statistics Netherlands can be sectioned into *input*, *throughput* and *output*. The projects within research themes follow this flow. The aim is to realise improvements in every step of this process.

In the remainder of this document the path is continued going from outside towards inside the organisation: the collaborations with universities and other centres of expertise and research institutes is briefly discussed in chapter 2, the optimal subdivision into research themes in chapter 3, and project governance and organisation within each theme, and the links with the priorities of the other divisions of Statistics Netherlands, in chapter 4. To conclude chapter 5 provides a more detailed description per theme of its methodological content and aims.

References

Ashton, K., 2009, That 'Internet of Things' Thing, In: *RFID Journal*, 22 July 2009

Bakker, B., 2013, verslag Werkgroep Meerjaren-Onderzoeksprogramma

Floridi, L., Taddeo, M., 2016, What is data ethics? In: Phil. Trans. R. Soc. A 374: 20160360. http://dx.doi.org/10.1098/rsta.2016.0360

Hulliger, B., R. Lehtonen, R. Münnich, J.-P. Ertz, D. Parvan, L. Jacquet, 2012, *Analysis of the future research needs for Official Statistics* (Luxembourg: Publications Office of the European Union)

Popper, K. , 1934, The Logic of Scientific Discovery (Logik der Forschung, English translation 1959)

Popper, K. , 1956, Realism and the aim of Science

2. Collaboration with universities and other centres of expertise

Within the Netherlands the institution of Statistics Netherlands has a high reputation as an independent and impartial centre of expertise and facilitator of social research. The research areas Statistics Netherlands covers concern both the statistical foundations and theory, responsibility for which lies with the teams of methodology, and the practical output and social questions for which it is essential that such methods have been applied appropriately and correctly by other teams. The result of this interweaving of methodological expertise in the entire organisation produces a steady stream of high quality and socially relevant publications.
The subject-specific and methodological expertise that Statistics Netherlands has at its disposal, together with the wealth of data that it collects and archives, make Statistics Netherlands **the** centre for knowledge transfer between universities and research centres on the one hand and policy makers and society on the other. In order to ensure that there remain good lines of communication between Statistics Netherlands and universities, a number of extraordinary professorships are employed by Statistics Netherlands who have a part-time position at various Dutch universities who also have the responsibility to form a bridge or conduit for knowledge as well as people. The professors currently in post within methodology are active in the following research themes:

- New observing techniques & Data collection: prof. dr. ir. B. Schouten (Utrecht U)
- Data integration: prof. dr. B. Bakker (VU Amsterdam) and prof. dr. T. de Waal (Tilburg U.)
- Big data, data mining & AI: prof. dr. P. Daas (TU/e Eindhoven)
- Statistical modelling: prof. dr. J. van de Brakel (Maastricht U.)

Past experience shows the beneficial and stimulating effect that these extraordinary professorships have for the research programme, which will remain the case. One of the reasons is that it enables a much faster transfer of knowledge concerning algorithms, methods and techniques from universities to Statistics Netherlands. For universities as well, there is an added value in these contacts. The application of new techniques to official statistics can give rise to interesting new lines of research. Also, masters' students and PhD students can learn a lot by interacting with the real life problems to which they apply their own research. The progress and advances that have been made within Statistics Netherlands' research programme have been sped up significantly by employing such students in work experience and internship initiatives. This means that relatively modest investments in terms of finances or supervision time translate to an increased offer of services and products for Statistics Netherlands as well as a more efficient production.

Not only are there collaborations with universities but also with other research centres and institutes for knowledge transfer to businesses, such as for instance TNO, and the planning offices, but also the Dutch Central bank (DNB) in subject

areas that are of mutual interest, and for which sharing knowledge and expertise is of benefit to all parties.
In addition to this there are also collaborative efforts with a broader range of parties such as the Jheronimus Academy of Data Science (JADS) and with diverse private parties, again with knowledge sharing as primary aim.

3. Subdivision into research themes

In order to carry out the research programme efficiently it is useful to subdivide and categorise the research into themes that fit the various areas of expertise of Statistics Netherlands. This enables allocating optimally the right people and right resources to projects. In many cases a given innovation project will require expertise from more than one area, but even then a categorisation is helpful for the management and steering of research efforts.

A categorization into research themes that fits well with both the expertise and the needs of the divisions of Statistics Netherlands is listed below, and also illustrated by figure 1:

- **New observing techniques & Data collection**
  This area concerns innovation of primary observation (e.g. surveys) and hybrid forms of observation and data collection such as the parallel use of various media in order to improve response. This also covers the use of administrative sources and register data and the use of data collected directly through e.g. telephone apps and sensors. In particular in the areas of sensor data (Internet of Things) and volatile data sources ('big data') improvement in systems is necessary, adapted to the volume and volatility of such data. Evidently there is an overlap with the research theme Big data, data mining & AI.
- **Big data, data mining & AI**
  New non-traditional sources of data often are not structured and may well consist of images, text or natural language instead of numbers and counts. This means specialised techniques are required to extract and analyse information from such data flows. Examples of some techniques are text mining and natural language processing, machine learning, and (interpretable) artificial intelligence. There is also an overlap with the research theme data integration in the sense that population units must be identified correctly in order to be able to establish whether and what information is missing.
- **Data integration**
  Central to this research theme are innovation and improvements in efficiency in order to merge and manage all data flows into Statistics Netherlands. The aim of this is to produce coherent and comprehensive statistical estimates of high quality. For every unit in the populations for which Statistics Netherlands publishes data, all available information needs to be retrievable with minimal delay.
- **Data security**
  Safeguarding data and data access is a process that needs continual improvements. One reason is that increasing amounts of data on individual population members are available as open data. Another reason is that, with time, ever larger amounts of computing power are more widely available. For collaborative projects and synergy with external partiers, such as universities and other public organisations one needs to guarantee that the data can be collected and exchanged safely, fully taking into account all risks of disclosure. (privacy preserving data sharing & analytics).

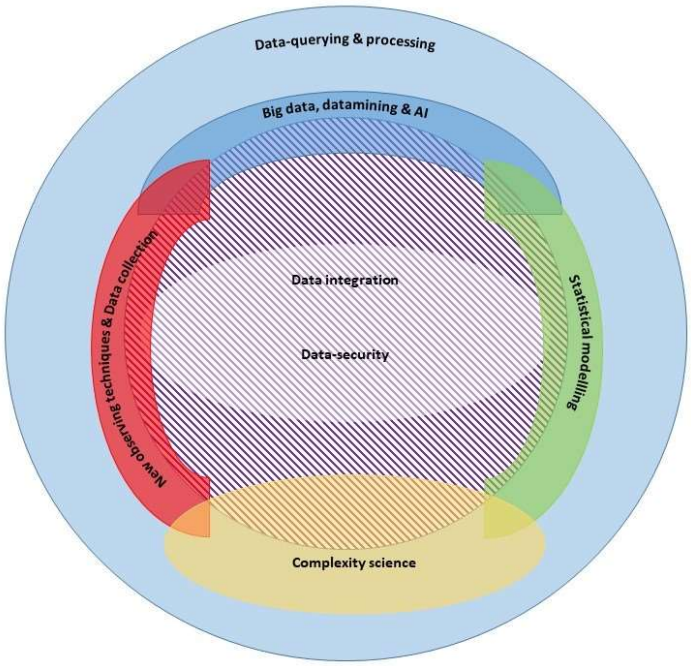

Figure 1. A diagram of the research themes in which the overlap of the various coloured areas symbolises that in research projects often the expertise from more than one theme is applied. The need for (innovation in) a good IT-infrastructure (both soft- and hardware) pervades all themes and is therefore shown by the light blue coloured circle in the background.

- **Statistical modelling**
  In order to satisfy the increased demand in high resolution geographical or detailed demographical information (complementary statistical services), it continues to be necessary to develop valid statistical models which provide solutions without bias and with minimal uncertainty. This concerns techniques for small area statistics, time series and nowcasting and quantification of trustworthiness and error. Bayesian techniques are to be used as supplement to the classical (frequentist) approach. The need for this gets stronger as more non-traditional data sources are being used.
  Also in the measurement of economic phenomena and in formulating which concepts need to be operationalised for measurement, for instance to construct index series, methodological techniques are used. In particular for new phenomena such as digitalization and the digital economy, more research is necessary.
- **Complexity science**
  Society consists of very many and very diverse actors and the relationships and interactions between them. Trends in society are in fact collective changes in this system of interrelationships. Positive and negative feedback, and non-linear mutual influences between various actors would normally be described or modelled at a 'microscopic' pairwise level. As a result this is expressed and measured at a macroscopic level in terms of emergent phenomena or trends. This is the field of complexity science. Statistics Netherlands is increasingly asked by policy makers to map out and quantify

mechanisms and relationships instead of just describing populations of people and companies. In order to continue providing answers in future it is necessary for Statistics Netherlands to apply the theories and analysis techniques of complexity science. This is to be used to clarify causal relationships in social and economic phenomena in society.

- **Data querying & processing**
  All the diverse types of data from a large variety of sources that are used must be processed rapidly, robustly, and stably. The processing and integration of data must lead to high quality statistics, including quality control measures to quantify that quality. These data continue to increase in volume, and the demands on processing and throughput speed also increase. This implies that both hard- and software must become better and faster as well. A central task is therefore good data management technology, capable of archiving, but also of querying, selecting, and filtering from the databases to obtain the right (sub-)populations and the right variables and enable answering specific statistical questions.
  In addition to this it is necessary to develop more versatile user interfaces for queries (also referred to as the information dialogue). To do this it is necessary to carry out and apply research into natural language processing & generation, and connected with this research the management of metadata for which a good metadata model is the foundation.

From experience of the past years it is clear that the majority of innovation projects within the research programme require expertise from more than one of the areas of expertise distinguished within the programme. Also with the new set of themes listed above, this will be the case. An example of overlap between research themes is between the themes “New observing techniques & Data collection” and “Big Data, Data Mining & AI”. The emphasis within the second theme is more strongly on the techniques required to manage and process a new or existing flow of data, in particular for non-traditional types of data. In the first theme on the other hand the emphasis lies on devising strategies to obtain appropriate data most efficiently. It is highly likely that some research projects to be taken up could be placed under either of these headings, and therefore in Figure 1 lie in the overlap between the areas coloured in red and in dark blue. The theme Data querying & processing has an impact on all other themes because for any research project there is a basic requirement that large amounts of data can be manipulated as quickly and transparently as possible. The themes “Big data, data-mining & AI” and “Data integration” overlap for instance wherever population units must be recognised and retrieved out of volatile datasets and combined with other more traditional data sources. The theme “Complexity science” partly makes use of techniques that are also used in the theme “Statistical modelling”, in different combinations or somewhat adapted for instance to map out relationships in networks or to validate models for the causal behaviour of dynamical systems. Methodologists of Statistics Netherlands are almost always active in more than one of the above themes.

4. Organisation of research within Statistics Netherlands

The organisational position of the teams of methodology within Statistics Netherlands is changing at the time of writing of this vision document. A directorate for Research & Development is being created within which the teams Methodology, the teams Process methodology and the Centre for Big Data Statistics (CBDS) are joined. This does emphasise much more strongly the importance of research and innovation for Statistics Netherlands.
One of the priorities and aims of this reorganization is that wherever possible the time span between methodology research via development to innovation and embedding in production is substantially shortened.

Methodological research can be seen as a process in which the first step is to devise fundamental methods, algorithms, and techniques which are validated for use in official statistics. In a second step such methods are made ready for production use, in close collaboration with senior statistical researchers throughout Statistics Netherlands. This means developing robust and versatile software packages and accompanying documentation, so that *state of the art* methods can then be implemented directly in primary processes. Finally there has to be a feedback mechanism in order to evaluate whether the latest methods and techniques continue to be appropriate given the changing needs of the various production divisions of Statistics Netherlands. There are regular contacts between methodologists and the research theme leads on the one hand, and statisticians in production teams on the other, for instance in the course of

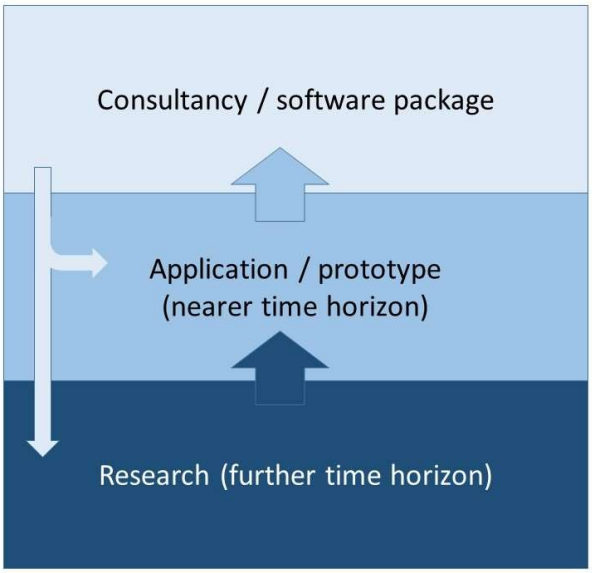

Figure 2. Within every research theme there are several types of projects. Some are more fundamental in character, while others are shorter turn-around research, which might also be classified as extended consultancy projects. In the end, each of these must lead to techniques and software that can be applied directly to any relevant official statistic produced by Statistics Netherlands.

consultancy projects or other ways in which methodology provides support and advice. Such occasions are used explicitly also to gather that feedback.

In general the research programme of methodology concerns itself with subjects with a time horizon of possibilities and needs that is further away than typical production time scales. Because of this, it can appear that the research has no obvious connection with the daily production processes of Statistics Netherlands. On the other hand, methods and techniques that were developed as part of the research programme in the past decade are now almost without exception widely used on a daily basis. Experience teaches that successful transfer of innovative methods depends on good communication and involvement between methodologists and researchers from other diverse production divisions.
In the previous period, 2015-2020, the research theme "Phenomenon-driven analysis and output innovation" was set up specifically to bridge the gap that had grown between methodological research and production. In the coming five years the experience from these efforts will be applied in all research themes to promote and facilitate technique transfer. That means that within every theme appropriate ways of working are to be adopted illustrated by figure 2. Within each research theme a number of specific projects are formulated every year. These projects are a mix of research with a longer time horizon, research directed towards creating documentation and software - minimally as *proof of concept* - and finally projects that are very close to the production of regular statistical output. In this last case the research is primarily to do with adjusting software modules and transferring knowledge of their use to production teams of the main Statistics Netherlands directorates EBN, SER, and DVZ.

Both in the directorate for socio-economic statistics SER (Chain effective use of registers, KERS) and the directorate for economic and business statistics EBN (EBN 2.x programme) there are broader innovation development programmes in place and being started up in 2020. It is inevitable that in the course of these innovations, new needs are going to be identified. It is organisational practice within Statistics Netherlands to designate the application of methodological expertise within such development projects as 'large consultancy'. Whenever the methodological aspects of innovation are directly, *off the shelf,* available, this is fully justified. However, past experience shows that in the course of such development tracks further questions of needs come to the surface for which there is no ready-made solution. For such cases the annual operationalisation of the research programme must be sufficiently flexible to take on board the necessary research projects. This feedback from the production of statistics on the research programme is also depicted in figure 2.

There cannot be flexibility in the research programme if the annual planning of staff capacity is set and controlled very tightly at the start of every year. While for the longer term projects, good resource planning is necessary to ensure continuity and the preservation of the knowledge base, some time must remain free and unallocated so that shorter-term projects can be treated according to *agile* business principles. At any moment it needs to be possible to add projects

to a backlog list of projects with an indication of the priority. Theme leads and methodologists must be given the space to re-plan their work in the light of these priorities. It is self-evident that both the researchers in production teams and research theme leads must contribute actively in removing any communication barriers that might exist.

In the past years, significant progress has been made in research into combining primary observation with adaptive strategies, together with registers, and volatile data sources. This has strengthened the position of Statistics Netherlands as the primary guardian and disseminator of factual information concerning Dutch society. The expertise and techniques which have been developed in this research are systematically incorporated into the processing stream at the level of *input*. That is a final good example of longer-term research which transferred to a *proof of concept* and is now in a phase where specific consultancy is available on demand as well as an internal offer of well-documented software-packages.

5. Detailed descriptions of the research themes

5.1. New observing techniques & Data collection

Communication forms, and the media that are used, between individuals, companies and governments have changed considerably. The introduction of all kinds of sensors and other forms of automated measurements have created new types of data. These trends will certainly continue in the coming years and represent great opportunities and challenges.

Both for observation of companies and of people, new forms of primary observation and hybrid forms of primary and secondary observation are considered in Statistics Netherlands' research. The interface with the theme Big data, data mining and AI is therefore large and cooperation with CBDS will increasingly be sought. New forms of primary observation concern both new observation instruments, such as mobile device apps, and measurements initiated by Statistics Netherlands via portable and fixed sensors, and so-called data donation where respondents themselves request and deliver existing data. Hybrid forms of perception arise when respondents are asked for permission to make links with existing (sensor) data. Various case studies with this hybrid approach are already being worked on for both companies and individuals. A number of them are well advanced and will be further elaborated in the coming years, including within Eurostat projects. The relationship with the DVZ's advanced data collection program is strong and will be used as much as possible.

These new forms of observation are especially interesting and promising if primary observation via questionnaires / diaries is very time-consuming, or if it concerns information that respondents do not know well, and/or if the concepts are difficult or impossible to put into operation via questions. This means that differences in statistics will arise compared to current observation due to the different measurements and/or concepts. This is not bad, both relevance and accuracy will increase, but it does mean that care must be taken with regard to comparability over time and between important target populations. The challenge lies in a good understanding of the characteristics of new observation and in a stable design of that observation.

To a certain extent, it is not yet known how willing companies and individuals are to contribute to various forms of observation. Research therefore focuses on effective recruitment and motivation strategies, attractive user interfaces and experience, a correct balance between active and passive measurements, efficient deployment of sensors, and a good understanding of ethical and privacy considerations. Target group approach, already a choice in primary observation, will become more important and need more attention. Planned missing designs for sensor measurements mean an extension of sample theory that must be worked out.

Hybrid observation and primary observation with sensors provides new data, often in a new, unknown form for Statistics Netherlands. Data science is an important discipline to derive desired output from existing survey themes. Here too, collaboration with the theme Big data, data mining and AI is crucial.

5.2. Big Data, Data Mining en Artificial Intelligence

The use of new sources, such as Big Data, for official statistics must become increasingly routine in the coming years. For this it is necessary to develop generic methods that make this possible. This includes methods for combining big data with other sources, methods for reliably extracting information from big data, and methods for correcting for the selectivity in big data. The diversity of new sources, and the large differences in quality within them, represent a major challenge. Major steps have already been taken in this regard in recent years, as demonstrated by the many Big Data-based beta products and the increase in use and interest in sources with other forms of data, such as texts and images, at Statistics Netherlands. This work will of course continue in the coming years with extra attention for extracting information about the units that generate the data in Big Data sources. This is necessary because such sources often contain hardly any directly available data about the units in the source and there is a clear need for this from the point of view of statistical production. In addition to continuing to build up knowledge in the Big Data field, the new research program also includes the most recent developments in the field of data mining and Artificial Intelligence (AI) which is used as a source of inspiration. The disadvantage of many of the most recent very successful approaches in data mining and AI is that they often rely on so-called model-based "black box" methods. As a result, it is not easy to determine how and why a certain method works so well. The clearest example of this is the success of neural network-based approaches such as Deep Learning and "word embedding" -based developments in Natural Language Processing. Although these approaches are very successful, it is very important for use in official statistics that the methods used are transparent, verifiable and fair. This therefore also includes the ethical component of the use of AI. These important additional characteristics are therefore explicitly included in the new research program.

5.3. Data integration

Official statistics are increasingly based on multiple sources. One can think of one or more surveys, administrative data, sensor data and unstructured sources such as internet data. There are all kinds of situations where sources are combined. This may involve the observations coming from multiple sources, but it may also be a situation where an additional source provides auxiliary information, for example to detect errors in a primary source. Although we aim for generic methods for all subjects, that is to say, widely applicable methods, they will not be able to cover all forms of data integration: there are too many different situations for that. In part, we will suffice with the development of practical tools for tackling any given problem.

The research theme focuses on the following four areas of attention (all four concern the context of combined sources and we provide a number of research questions):

- Linkage methods. How can we link files with different unit types? How do we optimize linking multiple files simultaneously? How do we link sources with custom-made products ("integration on demand") when the sources have little overlap or when there is a source (big data / sensor) without identifying variables?

- Editing and imputation methods. How far should we go with editing with integrated microdata with which different outputs are then made? How can we detect and correct errors in basic registrations using other sources? Can machine learning help make editing and imputing more effective? How can we improve on semi-automated machine learning?
- Estimation methods[1]. How can we set up "integration by design" where we start with already available sources (registers, online data, sensor data) and use surveys as a supplementary source? How can we correct for distorted level estimates or development estimates, for example by using an additional source? Can we monitor the causes, i.e. measurement and representation errors per subpopulation, as estimates of two sources deviate (too) far from each other; how can we reduce these differences or arrive at a combined estimate? How do we make clear estimates for developments and for estimates per domain, for output based on machine learning if the concepts differ over time and between domains?
- Methods for measuring (output) quality. How can we estimate the bias and variance in output based on multiple sources, with methods that can be used for multiple error types (coupling errors, classification errors, etc.)? How can we estimate the effect of multiple error types (e.g., sampling errors, register errors) on output quality?

5.4. Data security

Our care for confidentiality when handling data from individual units (individuals, companies, households, institutions, municipalities, etc.) is one of the conditions for a Statistics Netherlands' reputation of trustworthiness. It is not without reason that there are several articles in the government Act on Statistics Netherlands ("CBS wet") that deal with how Statistics Netherlands should treat data confidentially.

The media attention for the European GDPR that came into force in May 2019 (and its Dutch implementation, the AVG) has ensured that privacy has become a current issue again. Both data suppliers, research institutions and "the general public" are increasingly keen on privacy aspects.

Governments are putting more and more resources into making Open Data available. Collaboration should therefore become easier. However, the privacy of individual units is endangered by the combination of increasing amounts of data. Big data and other "new" sources of data are giving statistical bureaus the opportunity to publish phenomenon-oriented statistical information. But again privacy plays a role here: what can be found about (recognizable) individuals in the big mountain of data? How can you recognize individuals in a large pile of (unstructured) data at all?

Of course we have been paying attention to the privacy of our respondents since the birth of Statistics Netherlands. In particular since the end of the last century, a lot of methodology and software has been developed for statistical security: publishing statistical data in such a way that no information can be obtained from

[1] The focus area 'output quality' focuses on determining bias and variance for a given estimator based on multiple sources: that is, for output as it is currently being made. With estimations one tries to construct other estimators, namely such that bias is as small as possible (and often one also wants a small margin, or to be able to publish many details, or to be able to use existing sources).

(recognizable) individual units. However, this methodological and software-based toolbox is no longer sufficient in view of recent developments such as the four examples mentioned above.
Topics that are becoming important (again) include:

- Privacy Preserving Data Sharing (PPDS) and Privacy Preserving Analysis (PPA): collaboration between different parties to arrive at statistical information, without privacy sensitive information being (visibly) shared. Consider Homomorphic Encryption, Secret Sharing, Privacy Preserving Record Linkage.
- The speed with which publications become available, also based on big data / streaming data / sensor data / ..., places new demands on methods for statistical security. A frequently heard term is "confidentiality on the fly". It will have to be investigated whether this is really possible (privacy budget).
- Alternative forms of publication. Traditionally, the statistical information is first protected, after which a visualization is made. An alternative is to make a visualization based on unprotected (micro) data that must then be statistically protected. The loss of information in terms of visualization plays a role in this.
- The (renewed) interest in privacy aspects also requires a (renewed) view of the policy of Satistics Netherlands on privacy. Statistics Netherlands' starting position should be more based on explicit risk measures. This would mean that the risk would be quantified better and that the policy foundations would be laid down in an even better, more objective, manner.

5.5. Statistical modelling

The statistical modeling research theme focuses strongly on the use of new data sources that have not been obtained via probability samples. These are referred to as non-probability sample data and often also as big data. Broadly speaking, this type of data source can be used in two ways in the production of statistical information. The first option is to combine sample data with big data sources in model-based estimation methods, where sample data is the primary data source and big data sources are used as covariates. The second approach is to use big data sources as the primary data source.

The relevance of statistical information for users increases as figures become more detailed, with a higher frequency and more quickly available. Figures at municipality level or district and neighborhood level are more relevant than figures at national level. Figures on a monthly or quarterly basis are more relevant than figures on an annual basis. Figures that become available immediately after the reporting period are more relevant than figures that are published two or more reporting periods later. A problem with making detailed figures that relate to short reporting periods is that the available sample size is insufficient to make sufficiently reliable estimates using traditional methods from sample theory. A solution can be found in model-based estimation methods. This involves estimation methodologies that, through a statistical model, increase the effective sample size for the individual publication domains with information from other domains and reporting periods. These types of techniques are also particularly suitable for making more accurate initial estimates during the reporting period (nowcast).

Model-based estimation methods are very suitable for using new data sources such as big data in the production of official statistics in a relatively risk-averse manner, in particular because the Statistics Netherlands remains in control of the availability of the sample data that is primarily used for making publications . There are several complications to using big data as a primary data source that need to be taken into account. Firstly, these types of data sources may be biased for the target population to which results are generalized and the degree of bias is unknown so that it is difficult to correct for this. Further research into correction methods for bias in big data sources is therefore necessary.
To unlock new sources such as image and text sources, modern machine learning and AI techniques are necessary. Research into better understanding of how these techniques work and how they can be used responsibly in the production of official statistics is necessary. This indicates that there is overlap with the objectives within the expertise theme Big Data, Data Mining and Artificial Intelligence (section 5.2).
Finally, the use of big data sources as primary data source results in a risk increase in the production of official statistics because the Statistics Netherlands has no control over the availability of such sources and the comparability over time. That is why further research into the development of a responsible observation strategy with these types of sources is necessary. This means that there is close cooperation with the theme New observation techniques & Data collection (section 5.1).

Within this theme there are five sub-topics, each of which shows that there is a clear overlap in expertise and objectives with various other themes:

- Detailleerd estimation: There is a great and growing need for detailed estimates on a variety of topics. The desired detailing is sometimes regional, sometimes in time, and sometimes based on a combination of different background characteristics. The choice of model is very important for making detailed estimates. A number of guidelines already exist for this: the model must contain background variables that explain any selectivity of the data, and also the different levels at which it is estimated must be represented in the model. However, more research is needed into how these and other components can best be combined in a model and into ways to assess whether a model is good enough.
- Nowcasting and time series models: There is a substantial demand for timely and at the same time accurate statistical information. Nowcasting is a collective name for estimation methods with which provisional figures are produced in real time during the reporting period. Many of these methods use time series models, where a high-frequency auxiliary series is combined with a low-frequency series observed via a repeatedly performed probability sample. The relationship between the different series usually changes over time. Research into time series models that correctly model and estimate time dependent correlations is necessary.
- Estimation using non-probability sample data: It is necessary to explore the possibilities and develop estimation methods to use non-probability sample data in the production of official statistics. This is relevant because Statistics Netherlands wants to use Big Data, these data are often available for a selective subpopulation. Both when such data are used as an auxiliary resource, and also

when a statistic is fully based on big data, it is necessary to quantify bias and, where possible, correct for it.

- Deep learning and AI: New sources no longer consist of variables measured for population units, but can contain images and written and spoken text. Knowledge about deep learning is indispensable for unlocking image and text sources. Deep learning is a machine learning technique that consists of an artificial neural network with multiple hidden layers. Provided there is a lot of sample data, it is a powerful tool for learning complex relationships, deriving characteristics and predicting the value or category of previously seen examples.
- Observing strategies with new sources: It is difficult to meet new information requirements with just sample testing. Statistics Netherlands is already trying to optimize sample designs using available register information, but more and more data sources (such as new registers, sensor data, app information, etc.) are becoming available that could possibly be used for statistical output. A research question is whether and how these can be used to meet the information needs, as an independent source, in combination with a questionnaire or as help information with a questionnaire.
- Measurement of economic phenomena: Particularly in the case of new economic activities where there is no comparable existing activity, such as for example digital companies and the production and use of free services, measurement of economic quantities is complex. In most cases, a model must be used that correctly relates measurable quantities to quantities that are both socially relevant and well defined within economic theory. This is in part the field of economic research and falls under the AME research program (adequate measurement of the economy) set up by Statistics Netherlands' economic directorate EBN. However, this is a research area that also explicitly has a methodological component. Measuring the economy is also an active field of research at world-leading academic institutions, and it is necessary to develop and apply such expertise within Statistics Netherlands as well.

Comparison of economic variables in time or cross-section, i.e. between countries or regions within countries, is also theoretically and practically complex and there are not always obvious relationships between price, volume and productivity developments. This certainly applies to free services and digital products, but also to a number of services, such as medical care, where technological progress is strong. The definition of a product is not always clear here and the treatment of new and disappearing products, including quality changes, is a major problem. These issues also play a role in the processing of scanner data and online data in the CPI, where nowadays index formulas are known from cross-section comparisons and applied to time comparisons.

In addition, many social issues have dependencies over time, with choices now having far-reaching consequences in the future. Examples are the climate issue and the issue of the optimal use of scarce space and thus the functioning of the real estate market, including the housing market. It is necessary to develop adequate techniques for the corresponding comparison over time of economic and other quantities.

In practice, part of this work is taken up as large advisory projects, but there is an exploratory part that belongs to the research program.

5.6. Complexity science

This theme was new in the previous Long-term research program, so that the work still partly has the character of pioneering work. In short, Complexity science is about systems with interactions between the components for which the whole is more than the sum of the parts. That difference consists of "emergent properties" that arise from the interactions between the components and is by definition not entirely attributable to the behavior of those components per se. Complexity science has three ways to gain insight into these systems. Analysis of the interactions takes place with network analysis, analysis of the behavior of components is done with Agent Based Modeling and relationships between emergent properties are done with dynamic system analysis. In fact, these are three different perspectives to look at the same, and the results from one perspective can help to better understand them in another.
Completely in line with the ideas from the discussion paper "Application possibilities of Complexity Theory at Statistics Netherlands" from 2018, the emphasis in the recent period has been on experimenting with Agent Based Models and developing network data sets. There is now a first comprehensive data set of the Dutch people network based on register data, which is also being investigated. A method has been developed for attempting to create a dataset of the trade network between Dutch companies and tapping relevant sources. A first version of a dataset should be available by the end of 2019, probably for the greater part on the basis of model estimates.
Parallel to this, a relationship network has been established with a number of universities, policy researchers (such as EZK, TNO, CPB, DNB), companies (among which banks) and international parties (such as ENBES, some NSIs and foreign universities). In addition, some CBS members have become members of relevant communities (such as the Netherlands Platform Complex Systems and the Dutch Network Science Society). Both nationally and internationally, the responses to both network projects are very positive and there is much interest in collaboration.
The following steps are foreseen in the 2020-2025 period.

- Derive publications from the network data sets, including the development of methods to prevent data from being traceable to individual companies or individuals.
- The further development of the people network, among other things by making time series and adding new types of relationships.
- The further development of the method for creating a business network, tapping relevant sources and adding other types of relationships (eg between companies and employees and cross-border relationships of Dutch companies). In addition, it will be investigated to what extent the business network database can play a role in Statistics Netherlands' EBN 2.x project.
- Developing complexity applications for policy issues, in which Agent Based Models can play a role alongside network files.
- Depending on developments in network data and Agent Based Models, dynamic system analysis will also be taken up, preferably within the framework of policy issues.

With all these activities, the intention is to do this partly in collaboration with other parties, including universities, EZK, TNO and international parties. An

attempt will also be made to enter into a number of structural partnerships. There is already a strategic collaboration with the Institute for Advanced Study of the University of Amsterdam.

5.7. Data querying & processing

All themes in the research vision methodology are permeated with the use of modern information technology to make selections from data files and to edit or model that data. This is visually represented in Figure 1 where the theme "data querying and processing" has been placed as background and environment for all other themes.
At Statistics Netherlands, this research, which lies at the interface of IT and methods, has been placed organizationally in the IT development sector. IT research has three perspectives.
The first perspective fits most with the bottom layer in Figure 2. This perspective is that of IT as an innovator for Statistics Netherlands where the technology trends for the near future are assessed for their applicability in the Statistics Netherlands context and with a horizon of 2 up to 5 years ahead. The topics and priorities for this research work are determined by the IT department itself, whereby coordination with the CIO and cooperation with the CIO office is a matter of course. Collaboration with external parties is not shunned. This includes companies and institutions that are involved in Brightlands Smart Services Campus and colleges and universities.
Some concrete examples of topics that are being worked on are:

- the use of edge computing: where a number of data operations and selections are already done by the sensors or other data collection implementations, before the data is further processed in the internal computer systems of Statistics Netherlands. The technical possibilities of this and the implications are the subject of research
- quantum computing: this is a new technology that is likely to mature very quickly in the coming years. This can have major consequences for the privacy protection of our output, but also offers enormous opportunities in the field of data modeling.

The second perspective is that of IT as an enabler for the implementation of the Methodological Research and the (preparation of the) implementation of the results from this research. This is also explicitly stated in the vision of methodological research and collaboration between IT and Methodology is a matter of course. Prioritizing IT research from this perspective is mainly influenced by the ambitions of methodology.
Many of the strictest requirements for computing facilities are currently based on methodology. However, this automatically means that in a few years the same need has become necessary for the entire Statistics Netherlands when new methods are integrated into work processes.
When it comes to linking data, also with sensitive data from external parties, the solution does not necessarily lie with quantum computing. Processing large data streams is not the goal or the power of quantum computers. What is needed there is a better design (implementation) of data architecture and efficient algorithms and software. Examples are operations on large complex networks, as part of the

complexity research theme, and for "privacy preserving analytics" with external partners such as hospitals and for genetic research via, for example, ODISSEI. One can also think of applications for the "blockchain" technique. This perspective has a large overlap with the middle layer in Figure 2, but also the bottom layer.

The third perspective mainly belongs to the top layer in Figure 2. In this context, IT is a driver for the realization of the Strategic Agenda of Statistics Netherlands and then mainly focused on the ambition as expressed in #IT and #Innovation and the realization of the Grand Challenges (Big Data statistics, Advanced Data Collection and Information Dialogue). Methodological research also has this ambition, but there are also differences. The statistical departments are also innovating as a result of the pursuit of the goals from the Strategic Agenda and that will lead to the need for different use of IT facilities. For example, with regard to research into suitable models for metadata, the focus will be on its application for Information Dialogue-type applications. The issues involved include the automatic generation of derivation rules, the automatic generation of readable explanations and the general application of rewriting systems.
In part, work such as this can be incorporated into targeted advisory work, but it will also lead to research into the possibilities of IT that Statistics Netherlands is not (yet) familiar with. The prioritization of this research work follows from the prioritization of the Business Owners of the various innovation themes.

Due to the results of IT research in the 2017-2019 period, there is also international interest in the work of Statistics Netherlands in this area. In the period 2020-2025, international cooperation will be explicitly sought because, precisely when getting to know new and sometimes disruptive technologies, bundling research capacity will also lead to better implementation and acceptance of the results of that research. Collaboration is envisaged for the following forums: ESTAT / ESS, UNECE HLG MOS, UN Global Platform, UN Blue Skies Thinking Network, NITS, NITS + and bilateral contacts with IT colleagues from other NSIs.