STATISTICS ESTONIA

Development of environmental accounts

Activity 2. Developing environmental subsidies and transfer's account.

D1.3 Description of the methodology and methodological issues for environmental subsidies and transfer's account

Grant Agreement NUMBER — 101113157-2022-EE-EGD

Final Report

Raigo Rückenberg
Grete Luukas
Hans Hõrak
Peep Purje
Kersti Padu
Kaia Oras
Jane East
Marina Peiker

27.12.2024

Contents

We would like to thank Eurostat for providing the grant to develop Environmental Subsidies and Transfers Account and everyone contributing and consulting us on the topic. Special thanks go to Sjoerd Schenau and Marieke Rensman from Statistics Netherlands for consulting us throughout the grant project. We would also like to thank everyone involved from Statistics Estonia for their contributions and all the stakeholders for their inputs and feedback.

Introduction

This methodological report describes the development work done under Activity 2. "Developing environmental subsidies and transfers account" in the frame of the work on development of environmental accounts in under grant "Grant Agreement no 101113157 — 2022-EE-EGD, Development of the forestry, environmental subsidies and ecosystem accounts". We would like to thank Eurostat for providing the grant to develop Environmental Subsidies and Transfers Account.

Environmental subsidies or similar transfers are defined as current or capital transfer that is intended to support activities that protect the environment or reduce the use of and extraction of natural resources. All data within Environmental Subsidies and Similar Transfers (ESST) module are compatible with the concepts and definitions of the European System of Accounts (ESA) and System of Environmental Economics Accounts (SEEA).

This report will describe the data sources used, methodologies developed and summarize the results. In the annex, there are the minutes from the milestone seminars (involving stakeholders) and a summary from a study visit to Statistics Netherlands.

Previous experience with environmental subsidies and similar transfers in Estonia

During the previous grant project (101022852 — 2020-EE-ENVACC) methodology for compiling the ESST account was developed for the first time in Estonia. The previously developed methodology was used as a starting point for this grant project. This meant that during this grant project, a greater focus was placed on streamlining the processes and developing solutions for more efficient ESST compilation in the future.

However, several issues encountered during the previous grant project had to be addressed. As such, the compilation and development of the ESST account went on simultaneously and not all solutions could be applied directly in the compilation of 2022 ESST dataset.

In cooperation with Statistics Netherlands, several issues from the previous and this grant project was addressed:

- Determining flows involving local governments
- Integrating COFOG data into ESST account
- CEPA/CReMA classification of agricultural subsidies
- Agricultural subsidies that were assigned with wrong transfer type in the previous grant project were assigned correctly in this grant project
- Integration of Horizon and LIFE subsidies into ESST calculations
- Statistics Netherlands provided R scripts and examples of applications

Overview of the grant project

The objectives set in the grant agreement was to develop the Environmental Subsidies and Similar Transfers (ESST) account in Estonia and provide Eurostat with deliverables:

- D1.3 – Description of the methodology and methodological issues for environmental subsidies and transfers account for the year 2022
- D1.4 – Dataset on environmental subsidies and similar transfers for the year 2022

The main activities undertaken to improve the ESST account were:

- Analysis of previous methodology and data sources
- Signing contracts with data holders
- Classifying transfers according to ESST guidelines and ESA 2010
- Distinguishing Rest of the World (RoW, S.2) and General Government transfers (S.13)
- Analyzing and developing IT solutions for ESST
- Creating an Oracle database for ESST
- Cooperation with Statistics Netherlands to develop the ESST account in Statistics Estonia

All activities listed above were either started or successfully finished during this grant project. The activities still going on and carrying forward to the next grant project, require constant attention by nature. Such activities are signing and renewing contracts with data holders, and IT solutions, that are developed and updated in time. Despite this, greatest success in this grant project were various IT solutions developed. Specifically, application of R scripts, that instantly made an effect of the compilation of the ESST account in Estonia. Creating an Oracle database is still in its beta stage, but it was clear from the start that the creation of database for ESST was necessary, and it should be fully implemented for the next compilation of the ESST account. Another IT solution that shows great potential is applying machine learning to classify transfers according to environmental or resource management domain (CEPA or CReMA). The tool was trained and tested for CEPA/CReMA in this grant project and showed great potential, however, in the next grant project, the tool needs to be trained again for the new Classification of Environmental Purposes (CEP).

With new data sources available in this grant project, mainly Horizon/LIFE and Elering, more relevant transfers were integrated into the compilation of the ESST account compared to the previous grant project. Along with new data sources, the methodology was improved in a way that Rest of the World (RoW) and General Government transfers were distinguished more precisely – this significantly improved the accuracy of the results.

However, integration of the Public Sector Financial Statements (PSFS), so called COFOG data, remains problematic. Despite the discussions held with Statistics Netherlands (SN) and local National Accounts (NA) experts, it was not possible to integrate COFOG data fully into the compilation of the ESST account. Data from the EU was included in the compilation of the ESST account for the first time in this grant project, but the integration of the Horizon/LIFE projects poses certain issues, too. These issues will be addressed in the next grant project.

Determining the exact share of local governments in the ESST account remains an issue and will be further studied during the next grant project. However, it is assumed the local governments do not impact the final result of the ESST account significantly.

Additionally, the implementation CEP was discussed during this grant project. It is predicted that the implementation of CEP could cause some methodological and technical issues, but these issues will be further discussed during the implementation process in the next grant project.

In discussion with stakeholders, it was determined that tax abatements are not relevant in Estonia. However, the topic will be monitored and included in the ESST when it becomes relevant in the future.

Finally, the cooperation will continue between Statistics Estonia (SE), the stakeholders and SN. Input from the stakeholders and SN has proved to be extremely valuable during the last grant project. SE thanks both the stakeholders and SN for their vital input and looks forward to the cooperation in the future. Additionally, meetings took place between SE and Statistics Finland (SF) to discuss the compilation of the ESST account in both countries.

Results of the methodological development work on environmental subsidies and similar transfers (ESST) account are made available also on Statistics Estonia thematic website.

Development of the methodology and compiling of the account on environmental subsidies and similar transfers for the year 2022

The previous grant project was successful and the goal of compiling ESST account in Estonia for the first time, was achieved. However, during the development of the methodology for the ESST account, several issues were raised that needed further attention.

Solving the issues from the previous grant project will deliver more accurate results for the ESST account. Further development of the ESST account will help to streamline the workflow and reduce FTE's required to compile the ESST account in Estonia. As such, during this grant project, work continued to solve the remaining issues and develop solutions for automatization in the ESST account.

EU and General Government co-funding

It was observed during the last grant project that several support schemes include funding from Rest of the World (RoW) and General Government, so called co-financing. However, it was not possible to determine the share of funding coming from RoW and General Government. As in such cases, most of the funding originates mainly from RoW, and all the funding was assigned to RoW. This caused, in the previous grant project, the share of funding from RoW to be overestimated and funding from General Government to be underestimated.

To solve the issue, cooperation between SE and data holders was necessary to correctly identify such schemes. After such schemes were identified, contracts were made between SE and data holders to present the data in a way that RoW and General Government contributions could be identified correctly.

This issue has now been successfully solved. It is now possible to correctly assume the shares of funding to RoW and General Government, even when the support schemes use co-financing.

Transfers from General Government to local government

Determining and understanding the flows relevant to ESST from General Government to local governments proved to be problematic. It was understood that the General Government provides funds to local governments that are in scope of ESST. However, in most cases it seemed that local governments are not the final recipients of such transfers. The funds are in turn distributed by local governments to the final recipients. It proved difficult to identify such flows, especially when it came to RM activities, which are not covered in COFOG 05. Open data, published on local governments websites, are presented in a way that is unsuitable for data analysis.

To solve this problem, a number of local governments were contacted and data on transfers related to ESST were requested. Even when requesting data directly from local governments, the response rate was low, and it did not solve the problem of poor data quality. Still, the data provided by local governments was enough to make a rough estimate about how many ESST transfers are accounted for. To make the estimate, data provided by local governments were studied, along with administrative data. The estimates indicate that the share of such transfers in ESST is rather small – approximately 2% of General Government's contribution to ESST. As in the previous grant project, the transfers were excluded from ESST if it was not possible to determine the final recipient and the exact EP/RM activity of the transfer. Transfers from local governments will fully be included in the ESST once an efficient methodology is developed.

Overlap between Public Sector Financial Statements and other databases

During the previous grant project, it was observed that data from Public Sector Financial Statements (PSFS) overlap with the data from administrative sources. This observation was once again made in this grant project. Just as in previous grant project, PSFS (so called COFOG data) was excluded from ESST account, with the exception of D.74 transfers in COFOG 05 coming from Rest of the World. Since such transfers are made directly to the final recipient, it is assumed that such transfers are not shown in administrative data. Including other transfers from COFOG 05 would lead to double counting of a number of transfers. It was not possible to determine a pattern which transfers are double

counted – the transfers that appear one-to-one in both PSFS and administrative data seem random. There are also transfers where administrative data points to a final recipient, but PSFS data shows local government as the final recipient. Overlap between PSFS and administrative data remains an open issue and will once again be analyzed in the next grant project. Currently, there is no adequate solution on the table to solve the overlap/under coverage of COFOG and administrative data.

Aligning data between Public Sector Financial Statements and other databases

It was observed during the last grant project that transfers do not align between PSFS and administrative data. Compared to PSFS, administrative data has a lot more detail about transfers. The higher detail in administrative data allows for more precise classification of transfers. As such, the ESST account can't be completely aligned with PSFS data, and with National Accounts in general. From the identified overlapping transfers, it was observed that COFOG and classification based on administrative data were different. In some cases, this applies to transfer type, but more importantly, to the final recipients, too. Based on administrative data, it was possible to determine the correct institutional sector and NACE of the recipient. This was not possible using PSFS data due to the aggregation and detail in PSFS data.

Assigning institutional sector and NACE

During the last grant project there was a problem assigning the institutional sector and NACE to the final recipient. Linking ESST data with the business register for statistical purposes (compiled by Statistics Estonia) using MS Excel was not an optimal solution. In some cases, assigning the institutional sector and NACE had to be done manually and this caused some errors. Implementing R scripts meant that human errors were avoided in the process, such as typos and inserting data in the wrong data cells. This was confirmed by carrying out manual quality control to assure the script was running properly and that all the transfers were classified correctly according to institutional sector and NACE. Manual quality checks will remain in place for the future to assure the quality of the ESST account in the future.

Improving data quality for more accurate results

As the data gaps and problems with data quality were observed during the previous grant project, the filling of the data gaps and improving the data quality for ESST account was determined to be one of key issues to keep working on. The work continued during this grant project to improve data quality and fill the data gaps for the ESST account. The main focus was making contracts with data holders or updating contracts to match the needs of the ESST account. The process of putting new contracts in place or updating them is described in the "Data Acquisition" chapter. The work on updating contracts will continue through the next grant project, as the process takes time. The results of ESST account depend on the input data – the results can only be as good as the input.

Automatization for collecting and/or adding data to ESST account database

Another key issue from the previous project was to tackle the manual labor needed for the compilation of the ESST account. The datasets are large and working them through manually takes a lot of time. IT solutions and automated processes help to reduce the work done manually. Several IT solutions were analyzed and developed during this grant project. In the chapter "Analyzing the feasibility to develop IT solution and the process for future regular production of account" the IT and automatization solutions are described in more detail.

Creating a uniform database for ESST

When the work started in the previous grant project on compiling the ESST account for the first time in Estonia, it was unclear how the database for ESST account should look like or which software/environment to use. During this grant project, it was decided that Oracle database will be created for the ESST account. Such a database will ensure that the data is formatted uniformly and compiled in a single place according to the needs of ESST, with some processes being

automated already within the database, without any human intervention needed. More on creating a uniform database can be found in the subchapter "Creating an Oracle database for ESST transfers".

Data acquisition

As the compilation of ESST account in Estonia relies largely on administrative data, part of this grant project was dedicated to secure proper data flow to and within Statistics Estonia to compile the ESST account. During the previous grant project, data holders were identified and where possible, data was acquired via requests. This meant that the data acquisition relied on voluntary cooperation from data holders.

Although Statistics Estonia has good relations with data holders and data was successfully obtained to compile the ESST account during the previous grant project, it was clear that inquiring data on yearly basis and counting on the voluntary cooperation from data holders was not a sustainable solution for the future. Statistics Estonia would have no guarantees regarding data transmission and data holders would face unnecessary administrative burdens processing and answering Statistics Estonia data requests.

Statistics Estonia has a set of procedures in place to guarantee a sustainable dataflow from data holders, and within SE. For data holders, this meant legally binding contracts and obligations to provide data to SE, but also clear terms for data provision. Within SE, it means that all the data received was described and metadata attached to it. As such, projects were started in Statistics Estonia to acquire data according to the official procedures to guarantee a sustainable solution for data collection for the foreseeable future.

During this grant project, contracts were signed with KREDEX and ARIB. Projects to acquire data from State Shared Service Center (SFOS database) and Elering are still on-going and will be finished during the next grant project.

In addition to providing Statistics Estonia a fluent data flow, acquiring data according to the official procedures, it is possible to automize data cleaning and some data processing features for the compilation of ESST account.

Ordering Administrative Data

If the necessary data are not available at Statistics Estonia or are not with the required frequency, the missing administrative data must be ordered to Statistics Estonia. Before new data could be ordered, the leading analyst of the environmental statistics team had to make sure that the necessary data was not already available at Statistics Estonia. It was the task of environmental statistics leading analyst to determine from which institutions, from which datasets and which data fields should be ordered.

To process the request for new data, or make changes to the existing datasets, a project had to be created, described and approved first. The project description had to include its purpose, scope, legal basis, planned outcomes and impacts. Furthermore, the description had to address the problems the project aimed to solve and include any limitations or risks that might occur and affect the project. If there is no project or the project does not pass the internal review, SE will not contact data holders for any data.

Once the project was approved, a JIRA task (epic) along with subtasks was created to plan and monitor the progress of the project. All necessary communications and information were exchanged in JIRA, overseen by the leasing analyst from the environmental statistics team.

Projects to order data from Elering and SFOS have been approved and in various stages of their respective projects.

Generally, bringing data into the organization took at least 3 months after project approval, and in some cases, even 6 months if there were legal issues, particularly related to data protection.

Description of the Need for Requesting Administrative Data

To order the necessary data for ESST compilation, the leading analyst from environmental statistics team informed the coordinator from administrative data team about which institution or register and from whom (person or department) the data is needed.

The purpose and justification for requesting the data had to be provided. This included the name and number of the statistical work, citation to regulation 691/2011, but also the reasoning why it was not possible to compile ESST without requested data.

The scope of the request had to be described precisely by the leading analyst from the environmental statistics team. This included the frequency of data transmission, the observation period, the date of first data transmission, whether the data was needed retrospectively or for future periods only, fixed term or indefinite request or contract.

Along with the terms of the request, the exact data composition needs to be described by the leading analyst from the environmental statistics team. Conditions like description of variables in the dataset and data extraction conditions had to be submitted to the administrative data team designer. More precisely described data allows for easier negotiations with data holders.

Confirmation of Data Composition and Agreement

Once the data needs had been documented by the lead analyst of the environmental statistics team and the administrative data team had no further questions, the administrative data team contacted the data holders via email. In the email, SE explained the need for the data, from which data collection the data was needed, the required frequency, and the data composition. The data provider might have considered SE email as a clarification request and responded according to the law within 30 calendar days.

Once an agreement has been reached to receive sample data and the data has been sent to SE, for example, via email, it was loaded into the sample data source database (Final Observation Register – FOR) and, if necessary, anonymized - transformed so that it cannot be directly identified, i.e., pseudonymized. If the data was received via X-Road, it was parsed (transforming XML data into Oracle flat table format).

From the received data, SE verified whether the data composition was correct and all requested elements, variables and objects, were present; whether the received columns had the agreed-upon titles; whether the values of variables in the data matched the agreed lists (including spelling); whether the values met the requirements, including data types, etc.; and whether the data was available for linking and identification. The presentation formats of date type data were also checked.

Based on the analysis of the sample data, if needed, SE communicated with the data provider to improve data quality and then proceeded with formalizing the data transmission contract.

All the contracts signed between SE and data holders were reviewed by lawyers from both parties to ensure that all the legal basis were covered.

Creating an Oracle database for ESST transfers

Once the data has been obtained from data holders, it can be integrated into a database. From this database, the analysists in SE can view and download the data for processing. The benefits of having such a database are that all the metadata is described, and certain processes can be coded and automatically executed within the database.

The data storage infrastructure in Statistics Estonia is based on an Oracle database. Initially, raw data is extracted from various sources and transferred to the Data Staging Area (DSA) using the ETL tool. Within the DSA, comprehensive data transformations are executed, encompassing data cleaning, imputation, replacement and automated

computations. Post-transformation, the cleaned data undergoes anonymization and is subsequently loaded in a version-controlled format into the Final Observation Register (FOR) utilizing the ETL tool.

Our data pipeline utilizes the VAIS ETL tool, which was developed by Statistics Estonia in collaboration with external partners approximately 10 years ago. This tool is designed for loading, transforming, and validating data, and it meets the specific requirements of Statistics Estonia, including process and operations logging, as well as reusable templates to streamline the creation of data flows. Metadata management is handled by our metainformation system, iMeta.

During this grant project, we described data within our metainformation system (iMeta) and extracted subsidies data from various sources, including KREDEX, ARIB, EIC into the DSA using the ETL tool. Subsequently, we enriched the subsidies data with additional information, such as the field of activity, the number of employees etc., from the economic entities register. The versioned data was then loaded into the Final Observation Register (FOR). During this project, an automatic reusable workflow (which consists of VAIS ETL tool packages) was created to enhance efficiency and consistency in data processing.

Several challenges were encountered during the project. The analysis of data quality was time-intensive, and the data utilized was of suboptimal quality. Additionally, the relatively large size of the datasets resulted in prolonged processing times for the ETL tool.

Creating the database for ESST is still in its beta stage – preliminary version of the database was created, tested, and approved by a leading analyst in environmental statistics. However, during the next grant projected the ESST database will be further improved to fit the needs of ESST account better – this includes adding data fields, adding data from additional data sources to the ESST database and further automatization activities in terms of classifying ESST transfers, for example, assigning transfer codes automatically according to NA rules. Such a database could be a feasible solution for other countries, too, if large amounts of data from different data sources are being used for the compilation of ESST.

Methodology for the compilation of ESST account in Estonia

After analyzing the issues from the previous grant project and finding solutions where possible, improved methodology to compile the ESST account in Estonia was applied. In this chapter, the final methodology for this grant project is described. This chapter contains the description of:

- Data sources
- Methodology for classifying transfers according to environmental domain (CEPA/CReMA)
- Methodology for classifying transfers according to National Accounts (ESA 2010) classification i.e., subsidies, other current transfers and capital transfers
- Methodology for classifying transfers according to institutional sector of the final recipient
- Methodology for classifying transfers according to the industry (NACE) of the final recipient
- Integration to other environmental monetary accounts

Data sources

As mentioned earlier, the compilation of ESST account in Estonia relies largely on administrative data. The list of data sources with a short description attached to them follows:

- The State Shared Service Centre https://www.rtk.ee/en - SSSC database (SFOS) contains almost all subsidies paid by EU + some general government (S.13) transfers.

- The Agricultural Registers and Information Board (ARIB) https://www.pria.ee/en - deals with agriculture subsidies (from EU and S.13), such as organic farming (D.3 transfers) to investment grants (D.9), such as equipment to contain or reduce pollution from livestock, improving the energy efficiency of buildings, processes' and equipment, etc.
- Environmental Investment Center (EIC) https://www.kik.ee/en - most of EIC transfers can already be found in SSSC database, but some S.13 transfers are represented only in EIC database. Those are mostly transfers from S.13 to the final recipient. As the name suggests, all kinds of transfers are related directly to the environment.
- Enterprise and Innovation Foundation (KREDEX) https://kredex.ee/en - mostly related to investment grants, such as renovating buildings, building solar parks, etc. Funds distributed by this organization to the final recipient stem from S.13, no EU funds are dealt through this organization. As with EIC, most of the transfers are already found in SSSC data.
- ELERING https://www.elering.ee/en/renewable-energy-subsidy - State owned energy company that subsidizes renewable energy productions. Funding comes from S.13.
- Microdata from Local Governments – some larger municipalities and local governments provide us microdata about the subsidies handed out. This is a very small sum (around 2-3%) of all subsidies account. Usually, it is very difficult to obtain data from local governments/municipalities, since they are, too, overloaded with work and they do not gather such data to be used further or to provide anyone else. As such, the data quality is subpar, with very few exceptions.

- European Climate, Infrastructure and Environment Executive Agency (CINEA) https://cinea.ec.europa.eu/index_en - data for Horizon and LIFE projects. These transfers go straight to the final recipient, so they are not recorded in datasets mentioned above.
  - Public Sector Financial Statements (PSFS) - So called COFOG data. In Estonia we use only D.74 transfers originating from RoW. COFOG data is too aggregated, needs to be assessed and coefficients applied, and in general we call COFOG data "blind" or "dark" because of the lack of detail needed for ESST. In addition, when it comes to subsidies, we have a strong belief that all the transfers in COFOG data are already covered by sources above, in much more detail (except D.74).

The decision to use administrative data sources for the compilation of the ESST account was already made during the previous grant project. Some of the administrative data was already available, some administrative data had to be improved, and some had to be newly acquired. It was clear from the start, that administrative data in Estonia offers much more detail and is more accurate than PSFS data. Administrative data offers the possibility to separate transfers from RoW and General Government, as well as much more detail about the transfers and final recipients. In addition, it is possible to identify EP or RM activities from such support schemes that by default don't fall in the scope of the ESST account, but the technical description of the individual projects fit in the scope of the ESST account.

The administrative data available to SE also includes some transfers within general government (D.73). This has been confirmed by comparing PSFS (COFOG) data and administrative data. In addition, COFOG 05 in PSFS consists of CEPA domain, leaving CReMA domain uncounted for.

However, there could be some under coverage of other current transfers (D.7) with the extensive use of administrative data – administrative data contains mostly subsidies (D.3) and capital transfers (D.9). However, since it was not possible to determine the exact extent of overlap between PSFS and administrative data, PSFS data was mostly excluded to avoid double counting of transfers. The integration of PSFS data with administrative data will be further investigated in the next grant project with both Statistics Netherlands and SE's own National Accounts specialist.

Although the process of acquiring administrative can be time consuming at beginning, it is well worth it in the long run – the detail of administrative data is so much better compared to PSFS data, that it allows for more precise allocation of transfers and enables the use of IT solutions for the compilation of the ESST account in Estonia.

Methodology for classifying transfers according to CEPA/CReMA

Allocation of the environmental domain was done in line with Environmental Subsidies and Similar Transfers guideline and Classification of Environmental Protection Activities and Expenditure (CEPA) and Classification of Resource Management Activities (CReMA).

To allocate transfers into CEPA/CReMA categories, the support scheme and project description were analyzed. All the transfers were allocated a single CEPA/CReMA category. This was done so for several reasons:

- The support schemes in Estonia have a very narrow focus – so in most cases it was easy to determine the exact CEPA/CReMA category for projects under each scheme
- Projects descriptions are detailed enough to determine the CEPA/CReMA category
- In most cases it was possible to distinguish the main CEPA/CReMA activity from the project descriptions, if there were two or more CEPA/CReMA activities done
- Dividing transfers between several CEPA/CReMA categories would require a methodology to be developed and/or input from experts in specific fields of CEPA/CReMA
- Dividing transfers between several CEPA/CReMA categories would be too time consuming for the effect it has – such transfers make up for a low percentage of the total ESST account
- The only exceptions are activities related to wastewater management (CEPA 2) and management of water (CEPA 10). In most cases wastewater treatment and water management activities are done simultaneously. An expert opinion was asked on how to classify such transfers. It was suggested to use CEPA 2 60% and CReMA 10 40% shares
- If a project was mainly EP or RM focused and the non-EP or –RM activities proportion was small, the transfers was included fully in the ESST calculations
- If a project included EP or RM activity along with several non-EP or –RM activities, the transfer was excluded from the ESST calculations

For example, if a heat pump was installed and updates to the electric system were made to facilitate the installation of the heat pump, the transfer was fully counted as CREMA 13B.

If the electric system was upgraded in the whole building and a heat pump was installed in the process, the transfer was fully excluded from the ESST calculations.

Another example – if a building was fully renovated to improve the energy efficiency and solar panels were installed in the process, the transfer was fully counted as CREMA 13B.

If solar panels were installed and in the process the roof of the building was renovated, the transfer was fully counted as CREMA 13A.

As seen above, the allocation of CEPA/CREMA comes down to how the project was described and the subjective assessment of the person allocating the CEPA/CREMA category. Fortunately, the number of such borderline transfers was not significant.

Allocation of CEPA/CReMA categories manually takes up to three to four weeks. There were up to 50 thousand records that needed to be worked through for the year 2022, and it is likely that the number of records will not decrease in the foreseeable future.

The first step was to compare the observed period to previous years by linking the data with previous period. It is common that projects last for several years, and the payments are made over several years. When such projects are identified, they are assigned with the same CEPA/CReMA they had during the previous period. In addition to identifying several projects immediately, it ensures that projects fall in the same CEPA/CReMA category from year to year, thus improving the quality of the timeline.

The second step was to eliminate all the obvious transfers out of the scope of ESST. For example, support schemes for social welfare, non- EP or RM research and development schemes, infrastructure, etc. Projects within such schemes

were looked through to identify any keywords related to EP or RM activities. Some EP or RM projects were identified in the process, but most of the projects fell outside of the scope of the ESST account.

After that, schemes that fully fall under single CEPA/CReMA were allocated to the appropriate CEPA/CReMA. Examples of such schemes are Elering subsidies for production of renewable energy, schemes for improving the energy efficiency of buildings, subsidies for landscape protection and management. Again, technical descriptions were checked to make sure they match the predicted CEPA/CReMA category.

Finally, the support schemes with wider scopes were checked. Although within the scope of the ESST account, some support schemes cover projects belonging to various CEPA/CReMA categories. Working through such schemes took most of the time as the technical descriptions had to be studied for each individual project before they could be allocated to an appropriate CEPA/CReMA category.

Identifying which support schemes generally fall outside the scope of ESST, and which support schemes feature a heterogeneous mix of CEPA/CReMA projects, takes time and comes with experience. Fortunately, SE is working on developing machine learning tool that can do the same work almost instantly and manual input from the leading analyst is only needed for quality control.

Allocating PSFS COFOG 05 transfers to appropriate CEPA categories was done according to ESST guidelines.

Methodology for classifying transfers according to National Accounts (ESA 2010) classification

To classify transfers as subsidies (D.3), other current transfers (D.7) or capital transfers (D.9), ESST guidelines and ESA 2010 were followed. As in the previous grant project, no social contributions and benefits (D.6) transfers were found in Estonia that were in the scope of ESST. In addition, consultations with Statistics Netherlands and SE's National Accounts expert took place.

In some cases, the ESA transaction could be identified by studying the support scheme. For example, it was written in the description that a certain scheme is providing subsidies for certain activities or that the scheme provides funding for capital investments.

If it was not possible to determine the ESA transaction based on the support scheme, the owner, counterpart and account were studied, and classification was made based on said data.

If it was still unclear under which ESA transfer the transfer should be classified, the nature of the project was studied. Detailed project descriptions can provide a lot of information about where the transfer should be classified.

Classifying transfers was done in cooperation with SE's NA expert to ensure the maximum alignment possible between ESST and NA. As in previous grant, it was concluded that ESST has more detailed data available for classifying transfers, compared to NA, and more detailed data should be used to determine the ESA transaction. As such ESST and NA can't be completely aligned.

Cooperation with NA will continue during the next grant project to ensure the best possible alignment between ESST and NA.

Methodology for classifying transfers according to institutional sector of the final recipient

By using administrative data, it is almost always possible to determine the final recipient of the transfer. Administrative data also provides the name and business registry code of the final recipient. By linking the recipient's business registry code with the business register for statistical purposes it is possible to precisely determine the institutional sector of the recipient.

During this grant project, R script was written to make the linking process easier and faster. Business register for statistical purposes data was obtained straight from FOR and integrated into ESST dataset. This meant that there was no need to make MS Excel connections or manually link data between two massive datasets.

Methodology for classifying transfers according to the industry (NACE) of the final recipient

Classifying transfers according to the industry (NACE) follows the same logic as classification of transfers according to institutional sector. Data between ESST and business register for statistical purposes is linked and NACE is derived from the business register for statistical purposes.

Again, R script was written during this grant project to make the linking process easier and faster compared to previous grant project.

Integration to other environmental monetary accounts

ESST has a common part with EPEA and EGSS. ESST is an important data source for calculating output of some specific environmental goods and services. Data on environmental projects that have been subsidized are investments in use side, but in other hand it is production of output of environmental services and goods in supply side. EPEA uses output data on services related to CEPA categories, while EGSS uses data on output of environmental services and goods in CEPA and CREMA categories.

In some cases, ESST is the only data source for calculating the output of specific environmental goods and services as there is no other detailed monetary information that can be based on. These environmental services and goods are following:

- Protection of semi-natural landscapes (CEPA 6);
- Construction services for fish passages (CEPA 6);
- Transition of heating systems from fossil fuels to renewable energy (CREMA 13A);
- Renovating central heating systems (CREMA 13B);
- Energy efficient street lighting (CREMA 13B).

But there are also environmental goods and services, where several data sources for calculation of output are available and ESST is one of them. For example, output of service for cleanup of contaminated soil is calculated using investment data from statistical survey on environmental protection expenditures and data from ESST.

List of environmental goods and services, where the ESST is additional data source among other data sources, is follows:

- Construction of waste treatment facilities (CEPA 3);
- Cleanup of contaminated soil (CEPA 4);
    - Construction of noise barriers and non-motorized roads (CEPA 5);
    - Replenishment of fish stocks (CEPA 6);
    - Forest protection and regeneration (for forests under environmental protection restrictions CEPA 6 and for other forests CREMA 11);
- Energy efficient renovation (insulation of buildings) (CREMA 13B).

Data on environmental subsidies and similar transfers are gathered during the compilation of Estonian ESST, but also the data on self-financing costs of subsidized environmental projects were collected. Due to the detailed description

of environmental projects all relevant projects can be selected out and the total cost of these activities can be calculated, which also presents the size of the output of these specific services and goods in supply side (excluding value added tax).

ESST is very important data source for compiling EPEA as financing from and to rest of the world is one component in addition to consumption (current and capital consumption) of environmental services when estimating national expenditure on environmental protection. Still some differences have to be considered when using ESST:

- ESST includes CEPA and CREMA transfers and EPEA covers only CEPA;
- ESST classifies enterprises by NACE activity; EPEA does not have NACE activity for financing;
- ESST distinguishes ESA codes of subsidies and similar transfers, and this is not required in EPEA.

We see that results of ESST can be used directly for financing variables in EPEA as these accounts have significant connections, but it could not be done for 2022 compilation as analyses of final ESST results revealed that during the development of ESST some transfers were found that has not been considered in environmental expenditures.

Differences were found also in CEPA classification, for example support for environmentally friendly management is classified under CEPA 6 in ESST but the output or consumption has not been included to other EPEA variables under CEPA 6 and therefore financing cannot be subtracted. To ensure better integration between ESST and EPEA further analyses are necessary to see which subsidies and transfers are included to ESST and if these activities are also considered in EPEA. Statistics Estonia plans to extend the coverage of EPEA in 2025 by using additional information about subsidized activities from ESST next year when also changes due to CEP are analyzed and developed.

Analyzing the feasibility to develop IT solution and the process for future regular production of account

Part of this grant project was to analyze and develop IT solutions for the ESST compilation in Estonia. It is a logical step to develop and use IT solutions, as vast amounts of data are needed and processed during the compilation of ESST account. Lack of IT solutions became an acute problem during the last grant project when processing data and making calculations for ESST became increasingly difficult due to MS Excel not being able to handle the amount of data. Calculations slowed down, MS Excel kept crashing, and progress (data) was lost on a regular basis. This slowed down the workflow and made the compilation process longer than it should have been. Furthermore, working with a great number of data fields, human errors were to happen – data being entered in the wrong place or simple typos. This, in turn, caused errors in calculations that had to be manually identified and fixed. Developing IT solutions was seen as a solution to aforementioned problems.

Feasible IT solutions analyzed

In this grant project, several potential IT solutions were analyzed. Main IT solutions to analyze were using R for data analysis and developing a database for ESST transfers. Scripting and creating a uniform database are the most common IT solutions within Statistics Estonia, but other in countries also. In discussions within SE and with Statistics Netherlands, it was seen as an obvious path to develop such IT solutions during this grant project. In addition, during the project, machine learning and data mining also caught the attention of people involved in the project. As such, it was decided to analyze the feasibility of these IT solutions also.

Developing R scripts

The idea of developing R scripts for ESST was discussed already during the previous grant project. However, as it was the first attempt for SE to develop ESST, the focus was on identifying data sources and developing methodology for the compilation of ESST. During this grant project, it was possible to put focus on developing R scripts, too.

R scripts were developed and used to clean and process data for the ESST account. This made a huge difference both in the time and efficiency. From the analyst's point of view, the time saved by using scripts instead of doing manual labor in MS Excel amounted to almost three weeks.

During the data cleaning process, it was much easier to identify problems and format the data uniformly using R. Using R for cleaning the data was not a separate goal of this grant project, but nevertheless, the advantages of using R in such manner became obvious very fast.

R scripts were also used to integrate Horizon and LIFE projects data to Estonian ESST dataset. Scripts provided by Statistics Netherlands proved to be useful in this case. This highlights the possibility for reusing scripts and sharing information via scripts.

For this grant project, the main goals for using R were to identify the institutional sector and NACE of the final recipient. This goal was fully achieved. This process involved linking ESST data with the business register for statistical purposes. This not only benefited ESST, but also EGSS and EPEA, as parts of the ESST transfers are used as input for the compilation of these accounts.

Data mining for ESST

Statistics Estonia has earlier experience with data mining for other statistical accounts and Statistics Netherlands has experimented with data mining for EGSS, so the feasibility of data mining for ESST was also analyzed during this grant project. The aim was to fill the data gaps in transfers from local governments to the final recipient.

Local governments publish their budgets and expenditures on their websites. As such, obtaining data from local governments websites was thought to be the best way to fill the data gap. Manually searching, downloading and processing the data is not feasible, as it takes too much time due to the number of local government websites to be screened. The data on local governments websites is also heterogeneous – both in terms of content and formatting. Using data mining would have saved the manual work of converting and formatting data to fit the needs of ESST account.

The idea was discussed during a study visit to Statistics Netherlands with their experts and later with experts from experimental statistics department in Statistics Estonia. In the discussion with experts from experimental statistics it was decided that data mining would not be the optimal solution to gather data from local governments websites. The data is needed only once a year to compile ESST account. This means that if there are changes to local governments websites, the code would need to be fixed. Screening for errors in the code and fixing them would take a vast amount of time and would not provide any significant gain in work hours. The heterogenous data would also make the code more complex and currently there appears to be no efficient way to obtain, clean and convert different file types and formats to suit the needs of ESST account. If local governments start publishing data in a more standardized way on their webpages, data mining opportunities could be explored once again.

Machine learning

After all the relevant data is obtained and integrated into a single database, classification of transfers begins. The main classificatory for ESST account is Classification of Environmental Protection Activities and Expenditure (CEPA) and Classification of Resource Management Activities (CReMA). Nearly 50 000 transfers are analyzed and classified each year for ESST. Classifying such a number of transfers takes a lot of time when done manually. Even during the previous grant project, the idea of using some kind of IT solution to classify transfers according to CEPA/CReMA was discussed. Then, this idea did not gain much traction as experts from Statistics Netherlands or SE did not have previous experience with machine learning.

During this grant project, the idea was analyzed again in SE and a solution for classifying transfers according to CEPA/CReMA with machine learning was investigated. Machine learning showed a lot of potential for compiling ESST

account and will be worked on in the next grant project. It is possible to shorten the time of classifying transfers according to CEPA/CREMA, and in the future CEP, by approximately three weeks. Machine learning also pointed out transfers from previous periods where a wrong CEPA/CReMA classification had been assigned by human error and also highlighted transfers that are borderline – transfers which could be assigned to different or multiple categories. This also provided valuable feedback and points of discussion on how to classify such transfers in the future.

To obtain funding for any kind of project, usually a text document with a detailed description of the project has to be produced. If there is a need to group various projects based on their goals and activities, then these documents should carry rich information for classification. Instead of manually examining thousands of project descriptions to make classifications, machine learning can be leveraged to automate this process. This automation, however, still requires initial manual work to produce a training data set where examples of different texts are paired with the correct label for the project type (in this case the CEPA/CREMA category).

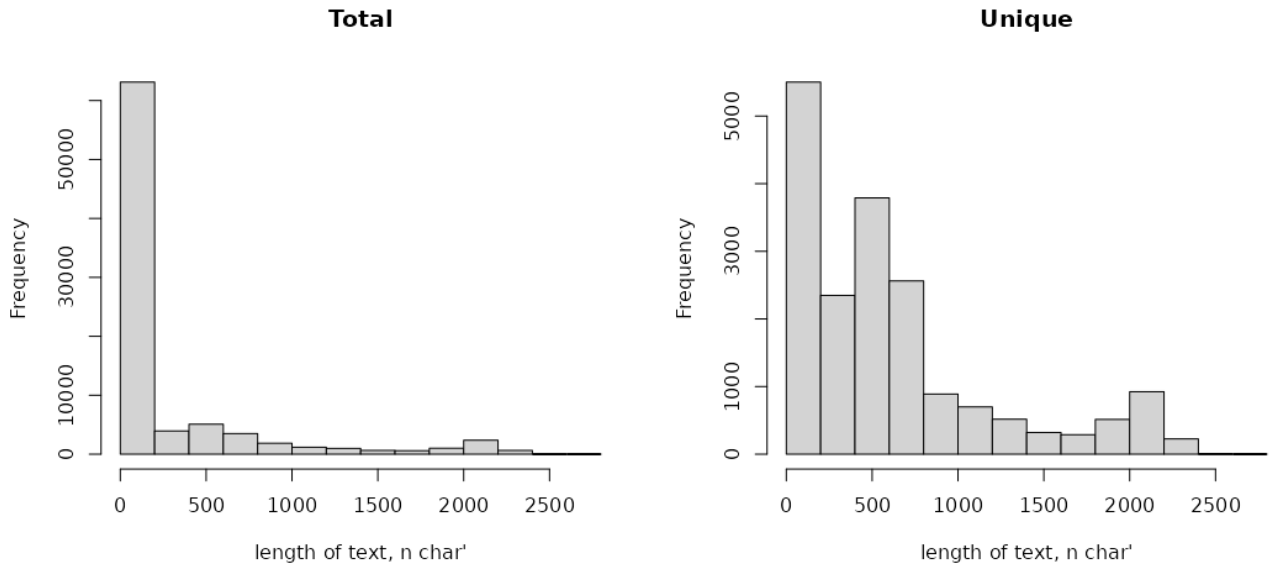
The training dataset is based on 84921 projects from 2020—2022 which were manually annotated with CEPA and/or CREMA categories or "other" (Table 1)To train and test a text classification model, we can only use unique text descriptions. Before selecting unique samples, all text fields available for each project were concatenated, punctuation was removed, and all text was set to lower case.

Tabel 1. Classification categories in the whole data set and data set with only unique texts

Category	Documents Total				Unique texts				class %, unique	% lost from dedup'	median text len', uniq'
	2020	2021	2022	Total	2020	2021	2022	Total unique			
CEPA 2/CREMA 10	1616	2012	589	**4217**	1514	1558	515	**3587**	19.3	14.9	534
CEPA 9	748	850	1070	**2668**	656	707	874	**2237**	12.04	16.2	165
CREMA 13B	488	631	1239	**2358**	231	192	306	**729**	3.92	69.1	329
CEPA 2	519	226	56	**801**	483	186	23	**692**	3.72	13.6	497
CEPA 6	2815	474	8408	**11697**	137	130	78	**345**	1.86	97.1	221
CEPA 3	206	61	73	**340**	51	31	60	**142**	0.76	58.2	396
CEPA 4	16501	46	14671	**31218**	92	28	17	**137**	0.74	99.6	278
CREMA 13A	62	95	72	**229**	50	24	42	**116**	0.62	49.3	453
CREMA 16	34	20	86	**140**	31	9	74	**114**	0.61	18.6	595
CEPA 8	36	53	63	**152**	29	33	49	**111**	0.6	27.0	202
CEPA 1	64	35	41	**140**	52	13	28	**93**	0.5	33.6	457
CEPA 5	32	53	14	**99**	31	24	8	**63**	0.34	36.4	906
CREMA 11B	29	67	10	**106**	29	20	4	**53**	0.29	50.0	1091
CREMA 10	21	12	4	**37**	21	12	3	**36**	0.19	2.7	387
CREMA 14	18	30	10	**58**	18	13	4	**35**	0.19	39.7	1247
CREMA 15	17	35	8	**60**	12	19	4	**35**	0.19	41.7	1059
CREMA 13C	15	20	7	**42**	14	8	5	**27**	0.15	35.7	861
CREMA 11A	6061	0	0	**6061**	9	0	0	**9**	0.05	99.9	553
CREMA 12	0	17	0	**17**	0	8	0	**8**	0.04	52.9	112
CREMA 11	0	7	0	**7**	0	6	0	**6**	0.03	14.3	89
CEPA 7	2	3	1	**6**	2	1	1	**4**	0.02	33.3	427
NOT CEPA/CREMA	6274	7236	10958	**24468**	3383	1974	4648	**10005**	53.84	59.1	538
ANY CEPA/CREMA	29284	4747	26422	**60453**	3462	3022	2095	**8579**	46.16	85.8	447
Total	35558	11983	37380	**84921**	6845	4996	6743	**18584**	100	78.1	477

Long project descriptions were available only for a subset of the projects which causes many duplicated texts: 66% of the initial data had text shorter than 100 characters (including spaces) which dropped to 9% after removing duplicated texts. Categories CEPA 4, CEPA 6 and CREMA 11a had mostly duplicated texts. After removing duplicates, dataset size was reduced by 78%. The distribution of text length in the whole dataset and for unique texts is displayed on (Figure 1).

Figure 1. Text length distribution in the initial data set and after removing duplicated texts



Aside from the loss of data due to missing full texts, another challenge for machine learning comes from class imbalance and presence of extremely rare categories: 11 of the 21 categories have fewer than 100 unique samples, 4 categories have fewer than 10. Even though text can carry very distinct class-specific information for training a generalizable model from only a few examples (e.g., the word "radioactive" should come up in and only in projects under category CEPA 7), assessment of model performance is severely limited if we have few test samples or none. For this reason, model performance on the rare categories is assessed by manually checking the model predictions for 2023.

The nature of the task allows a two-step classification approach:

- Does the text belong to any of the CEPA/CREMA categories (binary classification)?
- If it does, which CEPA/CREMA category is it?

The projects under the label "not CEPA or CREMA" can be very diverse – all sorts of activities in different fields of life except environmental protection. Some of these texts may include words which also appear in CEPA/CREMA projects. By taking the two-step approach, we can model this diverse group against any kind of CEPA/CREMA projects so that the CEPA/CREMA classes can be modeled against each other in a more homogenous feature space in the second step. The first model should learn general features of projects which are not aimed at the environment. The second model then only has to deal with projects specific to environmental subsidies – it can learn to distinguish CEPA/CREMA classes from each other without needing to separate each class from the "other" category as well.

We used the fastText algorithm (Bojanowski et al., 2017a; Bojanowski et al., 2017b) for machine learning. This algorithm represents texts numerically by combining word-level embeddings with subword information (character n-grams) such that a word can be represented as the sum of its character n-gram vectors. The method is extremely fast and the use of subword information allows good performance on rare words, making it especially useful for

classification tasks with several rare classes. Moreover, pre-trained word vectors (Grave et al., 2018) are available[1] for 157 languages including Estonian.

The speed of fastText makes the search of optimal model hyperparameters fast and easy. We started from the default settings provided in the fastText R library (Mouselimis, 2024), adjusted the learning rate and number of epochs over a few iterations, and then proceeded to find optimal character n-gram lengths. We found that a minimum of two and maximum of 5 characters (including characters signifying the beginning and end of words) give best results for the Estonian language at least in this specific classification task. The rest of the hyperparameters were determined iteratively by maximizing test set performance while avoiding overfit. If the model appears to overfit, we can reduce model dimension, number of buckets, learning rate and/or the number of training epochs (Table 2).

Table 1. Hyperparameters for fastText models of the two-step classification system.

Hyperparameter	In R code	Binary model	Multiclass model
Dimension of word vectors	`dim`	120	120
Number of buckets	`bucket`	1100000	1200000
Size of context window	`ws`	5	5
Maximum length of word n-grams	`wordNgrams`	2	2
Minimum length of character n-grams	`minn`	2	2
Maximum length of character n-grams	`maxn`	5	5
Minimum number of word occurrences	`minCount`	7	4
Learning rate	`lr`	0.19	0.24
Training epochs	`epoch`	11	14

The presence of several very rare classes restricts the `minCount` parameter of the multiclass model to a small value since the rare classes might only have a few occurrences of the words with crucial discriminative power.

Results (Table 3) indicate excellent performance of the binary model: almost 97% accuracy on a balanced test set with high precision and recall. The multiclass model achieves very good performance on some of the more common categories but there are several problematic classes as well. CEPA 4 (protection and remediation of soil and water) shows mediocre performance on the test set but takes up over a third of the initial data set – the model might not generalize very well even if the prediction year included long text descriptions. The category for general research and development (CEPA 8) seems to be a very heterogeneous group difficult to distinguish from other categories. On the test set, CEPA 8 is confused most often with CEPA 4 and CREMA 15 (Table 4**)**.

Table 2. Class representation and model performance for multiclass and binary models

Class	n training samples	n test samles	% in training data	% in total data	Pre-cision	Recall	F1-Score
CEPA 2/CREMA 10	951	249	24.8	5.0	0.919	0.956	0.937
CEPA 9	723	177	18.9	3.1	0.930	0.977	0.953
CREMA 13B	573	127	15.0	2.8	0.902	0.937	0.919
CEPA 2	513	137	13.4	0.9	0.889	0.818	0.852
CEPA 6	263	82	6.9	13.8	0.950	0.927	0.938
CEPA 3	125	17	3.3	0.4	0.636	0.824	0.718
CEPA 4	103	34	2.7	36.8	0.742	0.676	0.708
CREMA 13A	92	24	2.4	0.3	0.842	0.667	0.744
CREMA 16	94	20	2.5	0.2	0.857	0.600	0.706
CEPA 8	92	19	2.4	0.2	0.550	0.579	0.564
CEPA 1	74	19	1.9	0.2	0.824	0.737	0.778
CEPA 5	52	11	1.4	0.1	1.000	1.000	1.000
CREMA 11B	40	13	1.0	0.1	0.611	0.846	0.710

[1] https://fasttext.cc/docs/en/crawl-vectors.html

Class	n training samples	n test samles	% in training data	% in total data	Pre-cision	Recall	F1-Score
CREMA 10	30	6	0.8	0.0	1.000	0.500	0.667
CREMA 14	28	7	0.7	0.1	0.400	0.286	0.333
CREMA 15	28	7	0.7	0.1	0.286	0.286	0.286
CREMA 13C	24	3	0.6	0.0	1.000	0.333	0.500
CREMA 11A	9	0	0.2	7.1	-	-	-
CREMA 12	8	0	0.2	0.0	-	-	-
CREMA 11	6	0	0.2	0.0	-	-	-
CEPA 7	4	0	0.1	0.0	-	-	-
NOT CEPA/CREMA	7263	1260	50.1	28.8	0.961	0.970	0.965
ANY CEPA/CREMA	7226	1297	49.9	71.2	0.970	0.961	0.966
Macro score m17*					0.785	0.703	0.724

* Macro score across the 17 categories available in the multi-class model's test set

Table 3. Confusion matrix for the multiclass model.

True ↓ / Prediction →	CE 1	CE 2	CE2/CR10	CE 3	CE 4	CE 5	CE 6	CE 8	CE 9	CR 10	CR 11b	CR 13a	CR 13b	CR 13c	CR 14	CR 15	CR 16
CE 1	14								1			1	1				
CE 2	1	112	11		2												
CE 2/CR 10		21	238														
CE 3				14	2			1	1					2	1	1	
CE 4		1		1	23			3		2			1				
CE 5						11											
CE 6		1			1		76		1							1	
CE 8	1				4		1	11	1							2	
CE 9				1	1		5	1	173			1	3				1
CR 10										3							
CR 11b											11		1		3		3
CR 13a	1											16				1	1
CR 13b	2	2			1							6	119				2
CR 13c														1			
CR 14											1		1		2		1
CR 15				1				2		1			1			2	
CR 16											1				1		12

The use of pretrained word vectors had a major effect on the multiclass model but almost no effect on the binary model (Table 5). In this case, the availability of pretrained word vectors allows us to achieve performance required for practical use in automation. Without pretraining, the model performed very poorly on the rare classes. The uncertainty left in predictions for 2023 can be somewhat reduced by sorting the table with predictions by the size of funding so that the predictions for the largest projects can be manually reviewed. Additionally, the predicted class probabilities can be used to focus manual review on records where the model had lower confidence (potential borderline cases).

Table 4. Effect of using pre-trained Estonian word vectors on model performance.

Model		loss	Train accuracy	Test accuracy	Test Macro F1
Binary model	No Pretraining	0.126	0.972	0.966	0.966
	Pretrained	0.075	0.985	0.966	0.966
Multiclass model	No Pretraining	1.170	0.737	0.757	0.290*
	Pretrained	0.356	0.949	0.880	0.724*

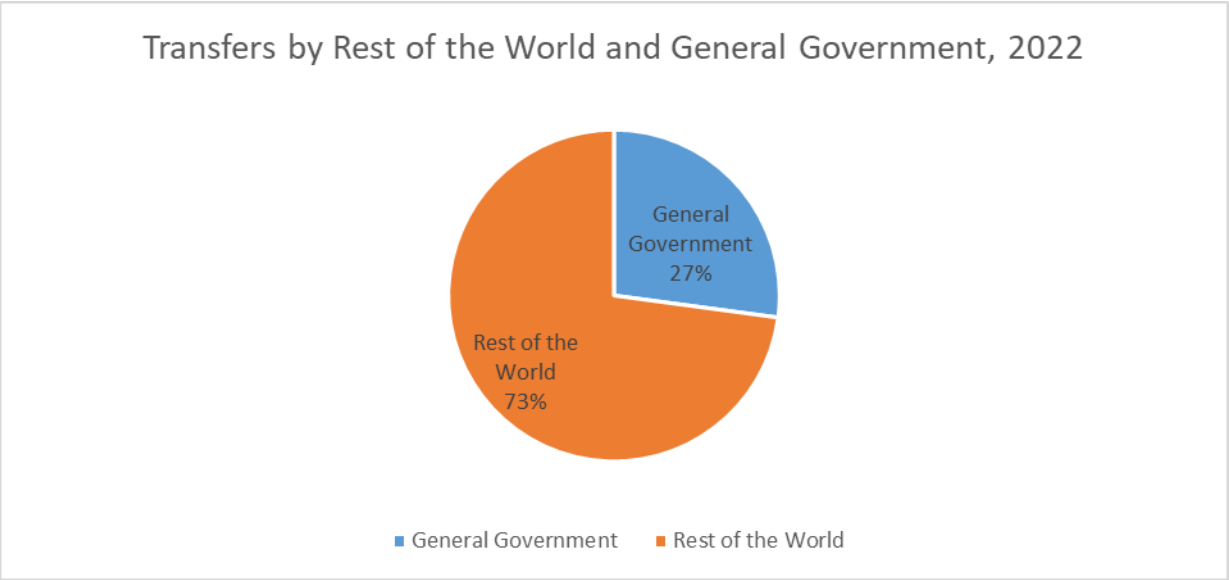
* Computed over the 17 classes available in test set

1 Results

Despite methodological and technical challenges encountered during this grant project, the complication of the ESST account can be viewed as a success. New data sources were included, and the quality of data was improved to make for more accurate results. IT solutions developed during this grant project made the compilation of the ESST account easier and faster and further developments should make the compilation even more streamlined.

As is previous grant project, it was observed that majority of the funding for EP- and RM activities originate from Rest of the World (Figure 2). Compared to previous grant project, the share of General Government has slightly increased, but this can be explained by including new data sources and improving the methodology to distinguish RoW and General Government funding more accurately. From the total of 505 million euros, contribution from RoW was 368 million euros and 137 million euros from General Government.

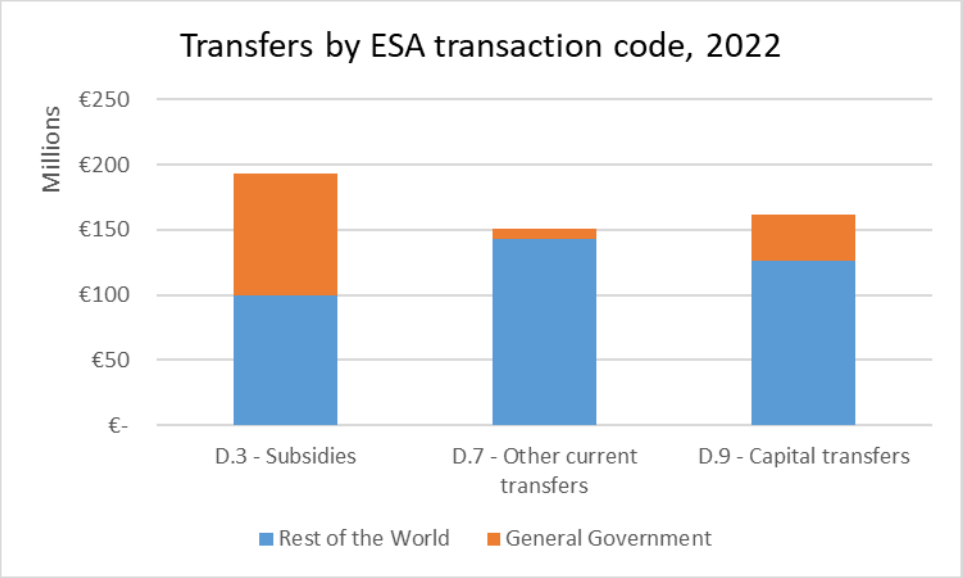
Figure 2. Transfers by Rest of the World and General Government, 2022



Transfers according to ESA transaction code were distributed fairly equally across the board (Figure 3). Subsidies (D.3) amounted to 193 million euros, capital transfers (D.9) 162 million euros and other current transfers (D.7) tallied up to 150 million euros (Figure 3).

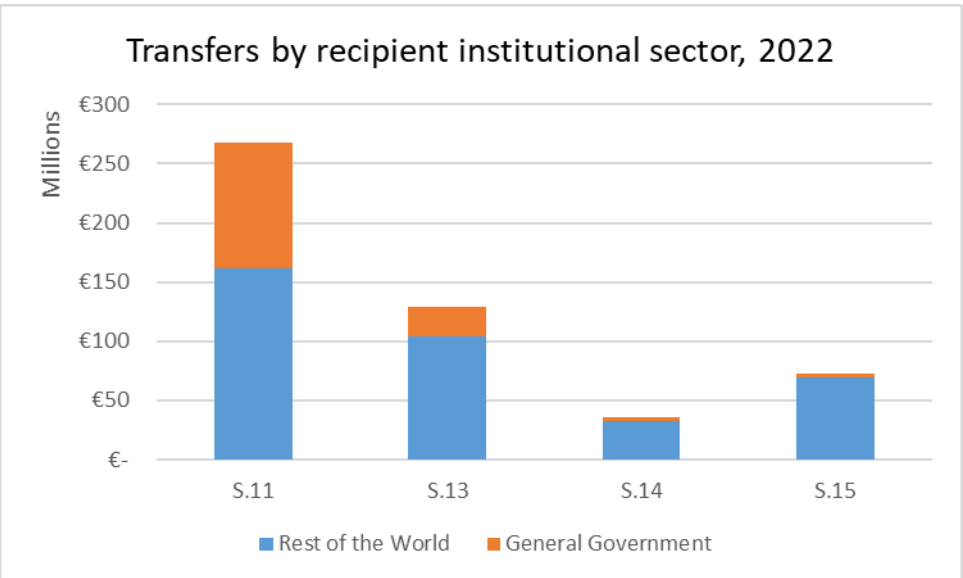
Subsidies (D.3) consisted mostly of subsidies for production of renewable energy from General Government and subsidies for sustainable agriculture from RoW. Other current transfers (D.7) consisted mainly of Horizon/LIFE projects for research and development. Capital transfers (D.9) were mainly made up of improving the energy efficiency of buildings and activities related to wastewater and clean water management.

Figure 3. Transfers by ESA transaction code, 2022



Corporations (S.11) are the largest recipient of environmental subsidies and other similar transfers in Estonia. Research and development (CEPA 8) activities made up for nearly half the funds transferred to General Government (S.13). These funds mostly originate from Horizon/LIFE programs. Non-profit institutions serving households (NPISH, S.15) received funds for heat and energy saving activities – improving the energy efficiency of buildings. Households (S.14) received the least amount of funds in total. Mostly these were transfers for organic farming, contributing to soil and groundwater protection (CEPA 4) and biodiversity (CEPA 6). (Figure 4)

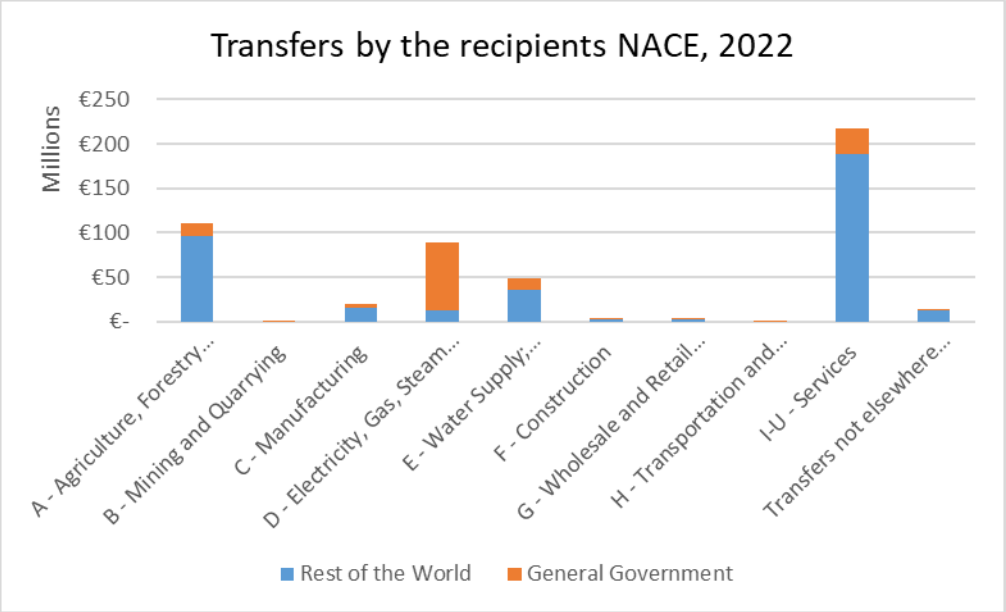
Figure 4. Transfers by recipient institutional sector, 2022



Highest amount of ESST transfers were received by the service sector (NACE I-U) (Figure 5). Economic areas such as Public Administration, Education and Real Estate were the main recipients in the service sector. Public Administration and Real Estate particularly received transfers related to energy efficiency and production of energy from renewable resources. Education received transfers related to research and development and Education, Training, Information provision and General Administration (ETIGA).
Another large recipient of ESST was Agriculture, Forestry and Fishing (NACE A). Much of the funding was related to subsidies for organic farming and measures for protection of biodiversity and landscapes.

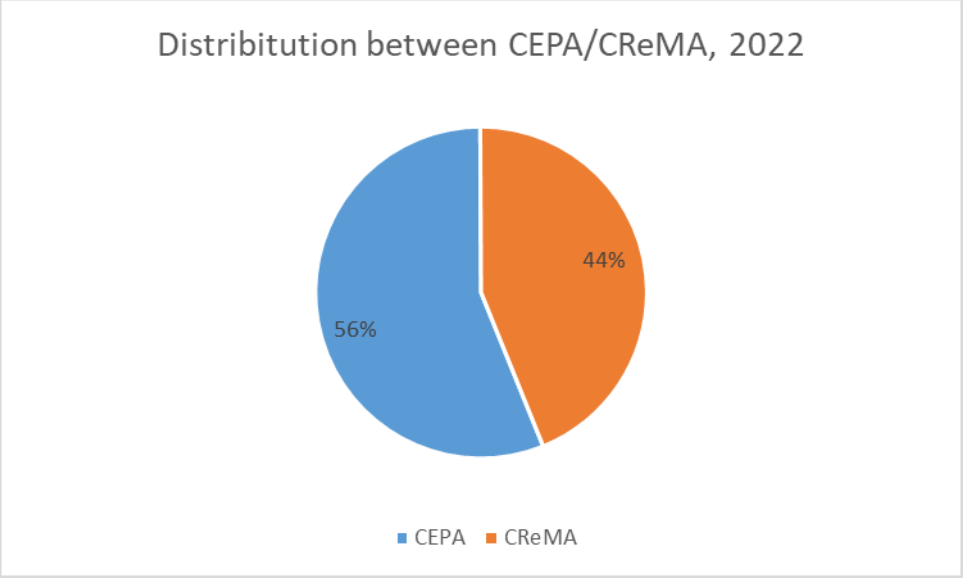
Electricity, Gas, Steam and Air Conditioning Supply (NACE D) also received substantial amount of subsidies for producing energy from renewable resources. Most of the funding originates from the General Government's scheme designated to producers of electricity from renewable resources.
Transfers not elsewhere classified (n.e.c) consist of households (S.14). Households mainly receive transfers for activities related to agriculture, production of renewable electricity and increasing the energy efficiency for buildings.

Figure 5. Transfers by the recipients NACE, 2022



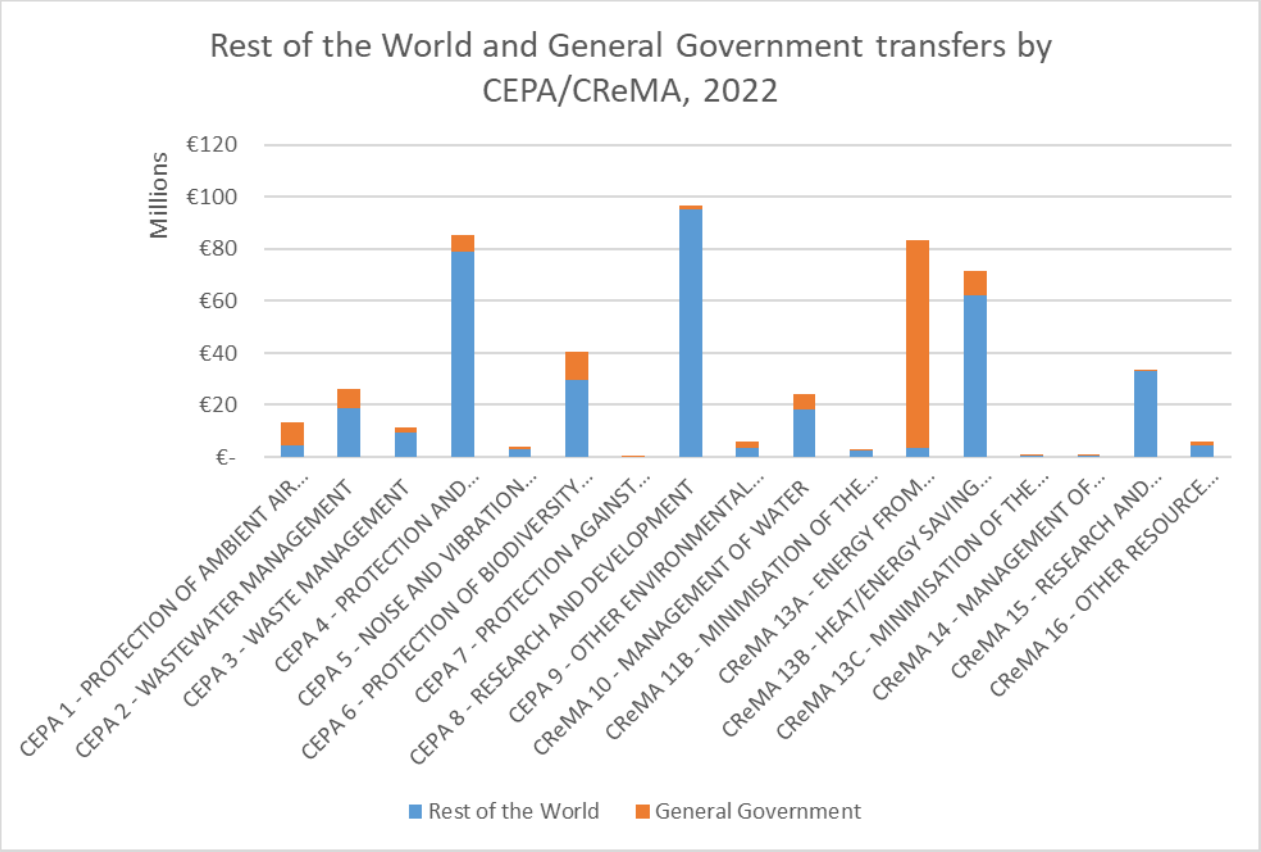
From the total of 505 million euros, 283 million euros, or 56%, was allocated for CEPA. For CReMA, 222 million euros, or 44%, was allocated (Figure 6).
Protection and remediation of soil and water (CEPA 4) and research and development (CEPA 8) received the largest amounts of funds in CEPA category. For CReMA, subsidies for the production of renewable energy (CReMA 13A) and heat/energy saving and management (CReMA 13B) contributed the most towards CReMA total.

Figure 6. Distribitution between CEPA/CReMA, 2022



Rest of the World and General Government environmental subsidies and similar transfers tallied up to 505 million euros in Estonia, 2022 (Figure 7). Research and development (CEPA 8) received the highest funding, 96 million euros, originating mostly from EU's Horizon/LIFE programs. Protection and remediation of soil and water (CEPA 4) received 85 million euros. These are mainly subsidies for organic farming. Once again, most of the funding originated from RoW. 83 million euros were contributed towards producing energy from renewable resources (CReMA 13A). However, in this case, almost all the funding originated from General Government, more specifically subsidies aimed at production of renewable electricity.

Figure 7. Rest of the World and General Government transfers by CEPA/CReMA, 2022



Remaining issues and further research

Although several issues from the last grant project have been solved, a few remained, and some new issues were encountered. Luckily, any of the remaining issues won't make the compilation of the ESST account in Estonia impossible. However, it is possible to improve the accuracy of the ESST account in Estonia. In addition, going forward with the development of IT solutions will make for easier and faster compilation of the ESST account.

The remaining issues and further research activities are described below.

Integrating Public Sector Financial Statements records into ESST

Using PSFS data for the compilation of the ESST account in Estonia proved to be an issue already during the previous grant project. An overlap between administrative data and PSFS data was observed. During this grant project the integration of PSFS data was studied further. With the help from National Accounts expert a better understanding of PSFS data was achieved.

However, it was not possible to decisively determine the pattern for the transfers that overlap or are missing from PSFS data. During this grant project, PSFS and administrative data were studied and compared manually. In the next grant project, trying to determine the pattern for overlapping or missing transfers should be studied. This would also enable the use of some automated processes to check for double counting.

Consultations with Statistics Netherlands should continue on this topic. It has already been determined that the data composition of PSFS data in Netherlands and Estonia are different. For example, Estonia has a set of account used by NA, that indicate if the transfer was for mediating foreign funds (accounts 450010, 450030, 450050, 450210, 450230, 450250). Studying both Netherland's and Estonia's PSFS data could offer some solution on how to integrate PSFS data into the ESST account in Estonia. It is recommended that National Account experts from both sides are available for consultations.

Integrating Horizon and LIFE projects into ESST

Horizon and LIFE projects were integrated into the ESST account for the first time in Estonia. However, not all projects could be included, specifically Cluster 6 projects.

Horizon and LIFE lump sum funds. In discussion with Statistics Netherlands, it was decided to follow their example and divide the financing equally across the project period (years).

Horizon and LIFE projects are available as open data, but the data composition makes it difficult to integrate them into ESST account. For example, the recipients names are spelled in many ways. The recipients names are spelled differently for different projects and business registry codes are not available. This makes it extremely time consuming to link the Horizon/LIFE projects with business register for statistical purposes and assign the correct institutional sector or NACE for the recipient.

During this grant project, the data was cleaned, and business registry codes were added manually to determine the final recipients institutional sector and NACE. This made the integration of Horizon and LIFE data a lengthy process.

Additionally, allocating Horizon/LIFE projects to a correct EP/RM category proved to be problematic. Approach similar to Statistics Netherlands was largely used – Horizon/LIFE projects were assigned to CEPA 8 or CREMA 15. However, when analyzing the projects descriptions, a number of projects descriptions were vague and abstract. Even if the support scheme was dedicated to R&D for EP/RM activities, the technical description did not offer a conclusive answer that the project fits the scope of ESST.

This raises a broader question: what should be prioritized when allocating CEPA/CReMA (CEP) - the policy/support scheme description or technical description of the project? To keep the ESST accounts comparable between different countries, a similar approach should be used in all countries.

There appears to be no overlap between Horizon/LIFE project data and administrative data. This is because Horizon/LIFE funds a distributed directly to the final recipient and therefore the transfers do not show up in any other database. This was also pointed out by stakeholders during the last grant project.

In cooperation with Statistics Netherlands, a better solution to integrate Horizon and LIFE data should be further developed. In addition, the open data published by European institutions should meet better quality standards.

Implementing The Classification of Environmental Purposes (CEP)

The biggest challenge of the next grant project will be implementing The Classification of Environmental Purposes (CEP). The new classification means both methodological and technical challenges.

The methodology for classifying transfers according to the environmental domain needs to be updated. As CEP is more detailed, it does not fully correspond to current CEPA/CReMA classification – implementing CEP will require studying technical descriptions of the projects even more thoroughly to allocate the transfers correctly. More specifically, allocating subcategories of CEP 01 (Air and climate) and CEP 05 (Soil, surface and groundwater, biodiversity and forest) could prove problematic as the project descriptions might not provide enough detail about the exact purpose of the transfers.

In some cases, it could mean that support schemes that were currently allocated into a single CEPA/CReMA category would now be split into multiple CEP (sub)categories. On the positive side, this means more accurate allocations of EP/RM activities, if indeed the technical descriptions of the projects are detailed enough.

From the technical point of view, some scripts need to be updated. However, switching to CEP will have the biggest impact on machine learning tool developed during this grant project. The tool needs to be trained again using the new classification. For the training process, previous years need to be revised and classified according to CEP. This means reclassifying approximately 100 thousand records.

Consultations with Statistics Netherlands will take place on implementing CEP.

References

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017a). Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2*, Short Papers, 427–431.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017b). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5, 135-146.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. arXiv preprint arXiv:1802.06893.

Mouselimis, L. (2024). fastText: Efficient Learning of Word Representations and Sentence Classification using R. *R package* version 1.0.4, https://CRAN.R-project.org/package=fastText.

ANNEX 1. Milestone meeting 1, summary: kick-off meeting with stakeholders, 10.10.2023

Participants: Raigo Rückenberg (Statistics Estonia), Kaia Oras (Statistics Estonia), Grete Luukas (Statistics Estonia), Kersti Padu (Statistics Estonia), Kristi Loit (Ministry of Climate), Aire Rihe (Ministry of Climate), Aivi Aolaid-Aas (Ministry of Climate), Kadri Kask (Ministry of Regional Affairs and Agriculture), Helene Eenlo (Ministry of Regional Affairs and Agriculture)

Statistics Estonia (SE) goal: introduce the new grant project (101113157-2022-EE-EGD) to the stakeholders.

Introduction: stakeholders were introduced to Regulation (EU) No 691/2011 and proposed changes to it – specifically making the ESST compulsory beginning 2025.

Results from previous grant project: SE gave an overview of the previous grant project (101022852 — 2020-EE-ENVACC), reviewing the methodology and results. Ministry of Climate (MoC) was interested in the visualization of the previous grant project results. Said materials were forwarded to MoC after the meeting.

Data and data sources: stakeholders were introduced with the data sources used in previous grant and potential new ones considered for this grant project. Stakeholders drew attention that SE has not included Horizon/LIFE projects in the compilation of ESST. SE will attempt to identify and include said projects in the ESST account in this grant project. The issue regarding local government subsidies was discussed – it was concluded that direct contact with local governments would be the best option to obtain reliable data from them. For other data sources, contracts have already been signed or are in the process of getting signed.

Issues with data: relying on administrative data poses a risk of some transfers being left unaccounted for, if the data holders or schemes are not correctly identified by SE. In addition, support schemes for EP/RM activities could change from year to year, further increasing the risk that some transfers could go unaccounted for in the future. Knowledge and input from stakeholders is needed to prevent such situations.

Development of IT solutions: one of the main goals of this grant project is to develop IT solutions to make the compilation of ESST account easier, faster, and more accurate. Stakeholders were introduced with SE plans to implement R scripts and creation of Oracle database for ESST. In addition, preliminary results of testing was introduced to stakeholders.

Discussion: MoC was interested if SE was planning to develop a database/app/portal for enterprises where they could check their contribution towards EP or RM activities (through investments or direct participation). SE has no such plan as there are not enough resources for that, also legal issues would arise with such project. Further, MoC was interested if SE evaluates the broader effect of environmental subsidies and similar transfers, eg., the value created per every euro spend for EP or RM activities or if SE gives any assessments regarding transfers made in scope of ESST. SE does not do such things, but ESST can be used as an input for policy makers.

Conclusions: cooperation will continue with stakeholders and a final seminar will be organized to introduce the results of the current grant project. Similarly, SE will continue cooperation with foreign experts (mainly Statistics Netherlands) to solve the issues related to the compilation of ESST and improve the methodology even further.

ANNEX 2. Milestone meeting 2, summary: methodological seminar I with Statistics Netherlands, 14.11.2024

Participants: Raigo Rückenberg (Statistics Estonia), Kaia Oras Grete Luukas (Statistics Estonia), Kersti Padu (Statistics Estonia), Sjoerd Schenau (Statistics Netherlands)

Statistics Estonia (SE) goal: giving SN an overview of progress made in the development of ESST account and finding solutions to methodological issues.

Introduction: Statistics Estonia (SE) is developing subsidies and similar transfers account in Estonia for reference year 2022. This is in accordance to proposed changes to regulation (EU) No 691/2011 which make ESST data submission mandatory starting from year 2025 (for reference year 2022). SE is currently working on improving the methodology for ESST account for reference years 2022 and 2023. In the previous grant (reference year 2020), methodology for compiling the ESST account was developed, however, some issues remained. As such, SE and Statistics Netherlands (SN) are working together to address these problems and finding potential solutions. The seminar held is part of the current grant project.

Presentation: during the presentation, SE gave an overview of progress made so far in the new grant project and introduced SE workflow in compiling the ESST account. The main points of progress are as follow:

- Addressing issues from previous grant project
- Contracts with data holders
- Preliminary testing of IT solutions for automatization
- Discussions with stakeholders – in Statistics Estonia (e.g., National Accounts) and outside of SE (e.g., Ministry of Climate)

Regarding SE workflows for ESST account, SN considered them to be very much alike to theirs, with the potential of developing IT solutions to fast-track certain processes.

Problems from the last grant project were then discussed. For some problems, a solution has been found. Some problems still need to be addressed to find the best solution. Problems from the last grant project are as follows:

- Transfers from General Government to local government
- Overlap between Public Sector Financial Statements and other databases
- Aligning data between Public Sector Financial Statements and other databases
- Assigning institutional sector and NACE
- Automatization for collecting and/or adding data to ESST account database
- Creating a uniform database for ESST

Transfers from General Government to local government – both SE and SN are facing the problem that transfers from General Government to local are governments are visible, however, the transfers from local governments to the final recipient are not detailed enough to identify whether it should be included in ESST account. One potential solution is web scraping the local governments webpages to find out if they support EP or RM related activities. For SE, it would be feasible to contact certain number of larger municipalities to find out if they support EP/RM related activities. Due to legislations, this is not possible for SN. Preliminary observations confirm, the local governments in Estonia support EP/RM activities, however, the scale of transfers is rather negligible, therefore a sensible solution must be found to identify transfers from local governments to final recipients – addressing all local governments with data requests is not a sensible to solution right now.

Overlap between Public Sector Financial Statements and other databases – this issue is solved by SE by substituting PSFS data for much more detailed data from other data sources. Most of COFOG_05 data available is found in other datasets. PSFS data should be used for transfers that do not show up in other datasets – transfers from local governments to the final recipient and transfers from General Government to Rest of the World.

Aligning data between Public Sector Financial Statements and other databases – this issue will remain until National Accounts changes their methodology. For ESST account, SE is using microdata that is much more detailed than

available to NA. Therefore, the methodology for assigning EP (RM not available in PSFS) activity and transaction codes is much different. This created differences is the previous grant project and will create differences between NA and ESST results in the future.

Assigning institutional sector and NACE – both activities are up for developing an IT solution. This will solve the problems from last grant project – human errors and typos. It will also speed up the whole process of compiling ESST account.

Automatization for collecting and/or adding data to ESST account database – this is solved by having contracts between SE and data holders. This will also speed up the process of compiling the ESST account. To develop an IT solution for automatization of certain processes, all data must be structured similarly and must be obtained according to SE data architecture. In the previous grant project, it was needed to contact each data holder individually, determine what data is available, how it is structured and how could SE gain access to their data.

Creating a uniform database for ESST – work has begun on creating a central database for ESST account. It is possible, that some parts of ESST data could be used for other environmental accounts (eg., EGSS, EPEA). In the previous grant project, a makeshift database was created in Excel. While it was possible to compile ESST account using Excel, it was found to be slow and inefficient.

Further, SE introduced the results of creating an IT solution for ESST. A visualization of previous grant project results was shared with SN. SE also shared the results of preliminary testing for creating an IT solution in R based on the last grant project data. SN shared their progress in developing IT solution and briefly introduced their data architecture. Regarding developing IT solutions, another seminar will be held between SE and SN in January 2024. Both parties will include their R specialist for more in-depth discussions on the topic.

Lastly, the influence of CEP on ESST account was discussed. The potential effects on moving forward with the new classification and revising previous years. When revising previous years, there is a threat of not having detailed enough data to classify transfers correctly according to CEP. For SN, it is up to decide whether to do revising year by year or all at once. This comes down to the resources available.

Conclusions: Due to the difference in data and methodologies, not all SN approaches are feasible for SE and *vice versa*. However, suggestion from SN to get in contact with limited number of local governments is feasible for SE. Similarly working together on IT solutions seems sensible – sharing the progress and problems that come up. A seminar focusing on IT solutions will be held in January 2024.

ANNEX 3. Summary: Study visit to Statistics Netherlands, 04.16.2024

Participants: Raigo Rückenberg (Statistics Estonia), Kaia Oras (Statistics Estonia), Grete Luukas (Statistics Estonia), Sjoerd Schenau (Statistics Netherlands), Julius Hage (Statistics Netherlands), Marieke Rensman (Statistics Netherlands)

Statistics Estonia (SE) goal: in cooperation with Statistics Netherlands (SN) find solutions to issues from previous and current grant project. Assess the progress SE has made in the current grant project.

Discussions

Data sources: SE uses mostly administrative data for the compilation of ESST. It was concluded that administrative data offers much more detail compared to COFOG data, which makes the allocation of transfers easier and possibly more accurate. SN does not have the administrative that to such extent relies mostly on NA data and has a very good cooperation with them. SE does work with NA to compile the ESST account but does not have such tight cooperation with NA historically.

In SN, COFOG transfers are assigned shares of CEPA/CREMA and updated regularly in cooperation between env. statistics and government statistics. Complex databases are set up and connected to allow for smooth updates and use in environmental statistics. Furthermore, SN can make suggestions to change the classification of certain transfers in public financing records.

SN uses data on spending from ministries to assign CEPA/CREMA shares to general government COFOG data. Transfers by local governments is still a problem – there is no central database for local governments transfers and it's difficult to acquire data directly from local governments and/or the data is not detailed enough. For this, 10 largest local governments are analyzed by SN to assume local governments spendings on CEPA/CREMA activities. Similar approach was used in SE to determine local governments spending on CEPA/CREMA activities – larger municipalities were approached with data requests and input. However, the response rate was low and data quality subpar.

SN advised SE that sometimes road maintenance could be classified under COFOG 05, although such activities fall out of ESST scope.

The financial transparency system (FTS) could be potential data source for RoW transfers. SN uses CINEA to determine the RoW transfers. This database could be used by SE to determine Horizon, LIFE, CEF transport/energy fund transfers, if they are not available from other data sources. However, the data is aggregated in CINEA and time of transfers can't be determined (the subsidy is accounted for the year project started). For more information, like the focus of programs, it is necessary to read the description of all programs and assume the main environmental focus and transfer codes. Horizon/LIFE projects were also brought to SE's attention in the kick-off meeting with stakeholders.

Scope of ESST: discussion were held on the topic of which transfers to include in ESST compilation. This issue is especially relevant for agricultural subsidies and the CEPA/CREMA classification of transfers should be further discussed as the schemes itself are similar across EU countries. For example, transfers that are classified as CEPA 4 by SE, are classified as CEPA 6/CREMA 13 by SN. This will make the results incomparable between countries.

For renewable energy production, SE should check for any possible tax abatements (netting schemes) available for the producers – right now SE is unaware of such abatements.

Further discussions should be held how to classify transfers coherently in ESST account in different countries. This applies to data for Horizon, LIFE, and other programs, but especially for subsidy schemes for agriculture. This is to assure that ESST transfers and CEPA/CREMA (CEP) categories remains comparable between countries.

Application of CEP: technical challenges, ideas were discussed. Main issue is that it is not possible to convert CEPA/CREMA directly to CEP for all categories. This means some challenges could lay ahead in regards to the methodology and coding. It was agreed that discussions regarding CEP should continue and be further discussed when CEP is applied.

IT solutions: application of several IT solutions was discussed. The main topic was applying R for the compilation of ESST in Estonia. SN was willing to share their R codes for various tasks. SN presented their experience with web scarping, something that could be explored in SE. SE presented their idea on machine learning to assign CEPA/CReMA to transfers. This idea received positive feedback from SN and will be explored further by SE.

ANNEX 4. Milestone meeting 3, summary: methodological seminar with Statistics Netherlands, 03.12.2024

Participants: Raigo Rückenberg (Statistics Estonia), Kaia Oras (Statistics Estonia), Grete Luukas (Statistics Estonia), Kersti Padu (Statistics Estonia), Hans Hõrak (Statistics Estonia), Taavi Dubinin (Statistics Estonia), Sjoerd Schenau (Statistics Netherlands), Marieke Rensman (Statistics Netherlands)

Statistics Estonia (SE) goal: introducing the results of the grant project. Describe the methodology and discuss issues and further development on ESST account in Estonia.

Discussion:

Results: SE presented the results of the current grant project. The results are not directly comparable to the previous period because of the improvements made to the methodology and inclusion of new data sources. SN has given advice to SE during the grant project on how to improve the methodology.

Methodology: In this grant project, SE has included Horizon/LIFE projects to the compilation of ESST. SN provided help and advice on how to include them and which schemes to look at. SE corrected the agriculture subsidies ESA transfers code classification after discussions with SN – in the previous grant the transfers were wrongly classified as D.7, and not as D.3, as they should be.

In addition, SE and SN propose to discuss with Eurostat how to allocate agriculture subsidies CEP in the future. The purpose of the funding from EU is the same or very similar for each EU member state, so it is only logical that in ESST they are allocated the same ESA transaction code and CEP category in the future. Currently, SE and SN have a different CEPA/CReMA allocation of such schemes. This makes the data not comparable between countries. It would be good to have a guideline or "rule of thumb" that all member states can follow on how to allocate EU agriculture subsidies.

SE gave a presentation on the topic of machine learning – the methodology and results. The results of machine learning were deemed a success as it was accurate enough and will make the process of allocating transfers to correct CEPA/CReMA category faster. However, in the next grant project the machine needs to be trained again according to CEP.

The integration with other accounts was also discussed – in SE both EGSS and EPEA use data from ESST database. For certain inputs, the ESST database is the only possible data source for EGSS and EPEA.

Issues and further research: integrating COFOG data to the compilation of ESST account was still problematic for SE this grant project. The methodology and options were discussed repeatedly with SN during this grant project. Also NA specialist from SE was advising environmental statistics team on COFOG issues. Unfortunately, the issues that prevailed during last grant project were still present in this grant project, too. Despite best efforts, the double counting and aggregation of COFOG data could not be overcome. It is acknowledged that COFOG data remains problematic to integrate into the ESST account in SE due to overlaps with other data sources. Work will continue in the next grant project to explore further options on how to integrate COFOG data into the ESST account in Estonia.

The cooperation between SE and SN will continue in the next grant project – better integration of COFOG data and Horizon/LIFE projects, switching to CEP, integrating ESST data to EGSS/EPEA and other current topics that arise during the development of the methodology for ESST in Estonia.