Journal of Economic Structures

RESEARCH

Open Access

What does e-invoice data bring to SNA and real-time economy?

Kaya Akagi[1*]

*Correspondence:
akagi_kaya@icloud.com

[1] Platform for Arts and Science, Chiba University of Commerce, Chiba, Japan

Abstract

This paper explores the potential of electronic invoice data to enhance the System of National Accounts (SNA) and facilitate a Real-Time Economy (RTE). The rise of data science and advancements in information and communication technologies have increased the use of big data in economic statistics. E-invoice data is being digitized in Europe to detect VAT fraud, but it also contains a vast amount of information on economic entities that make up the SNA. This study focuses on integrating electronic accounting information to improve the coverage, granularity, and update speed of SNA. The paper examines the concept of RTE, which emphasizes real-time data processing and automation, particularly in Estonia and Finland. It also highlights the benefits and challenges of using invoice data for economic statistics, including issues related to data storage formats, legislative systems, and the coordination of data linkages. The research demonstrates invoice data's theoretical and practical applications for constructing Input–Output Tables (IOTs) and monitoring financial situations. Despite the technical and institutional hurdles, the study suggests integrating invoice data can significantly transform economic statistics.

Keywords: Invoice, SNA, Input–Output Tables, Accounting data, Real-Time Economy, Exchange Algebra

1 Background

The rise of data science has significantly increased the utilization of various big data in national statistics. Advancements in information and communication technologies, coupled with the growing digitization of accounting data, have propelled discussions on integrating these data into economic statistics, particularly within the System of National Accounts (SNA). Doug Laney, an industrial analyst, coined the term "Big Data" to describe datasets characterized by three main attributes(3V): Volume (more than 1PB), Velocity(real-time), and Variety (e.g., formless text, video, and transaction data) (Laney 2001). In the past, the primary purpose of using big data in economic statistics was to improve data coverage (Variety) with unconventional formats such as tax and administrative record information.

However, in recent years, the focus has shifted towards the potential of big data to provide real-time insights (Velocity). Big Data can offer more frequent updates and real-time insights, a stark contrast to traditional survey methods, which are often periodic and slower to report. The COVID-19 pandemic has further accelerated the adoption and exploration of big data sources for economic statistics (Mehta et al. 2021; Bernal and Sejersen 2021).

In a highly digitized society, every economic transaction generates and processes electronic information to process and records it. Theoretically, utilizing such information makes it possible to instantly and automatically collect and compile information on all economic transactions. This paper explores the potential of improving the coverage, granularity, and update speed by applying electronic accounting information to SNA.

1.1 Real-Time Economy

Recently, the concept of a Real-Time Economy (RTE) is attracting attention in Western countries, especially in Baltic countries such as Estonia, to utilize big data for economic statistics, analysis models and Evidence-Based Policy Making (EBPM).[1]

The concept of current RTE was first proposed as a vision of General Electric in 2002, as a way to manage companies in real time through computerization and society as a whole to respond to changes in real time (Siegele 2024). In this sense, the concept of the RTE initially spread in the technological field of monitoring through the computerization of procedures.

On the other hand, RTE has also been a prominent research target in accounting, especially in auditing (Vasarhelyi 2023). In accounting, the Rutgers Accounting Research Center at Rutgers Business School started studying RTE in managerial accounting around 2002. The center describes the RTE as “The core objective of the real-time economy is the reduction of latency between and within processes.” (RARC 2023). The Rutgers Business School has been not used the term RTE itself before 2002. Still, they have used the concepts of continuous online monitoring of transactions, continuous online auditing, real-time auditing, and real-time reporting since before 2002 (Kogan et al. 1990; Victoria 2014).

As shown above, RTE started as a concept about managing companies using information and communication technology and developed as a new auditing technology in accounting. On the other hand, in recent years, RTE has been expanding its control target from a single company to national units. Many countries, especially in Baltic countries such as Estonia and Finland, focus on the RTE on the whole economy. Research on RTE covers the entire country in Finland starting around 2006 at the Aalto University School of Business. Since 2015, the Real-Time Economy Competence Center has been the core of the research.

Similarly, in Estonia, research is being conducted at Tallinn University of Technology. A survey study conducted by Tallinn University on behalf of the Estonian Ministry of Economic Affairs and Communications (MoEAC) describes RTE’s current status in Estonia (Robert 2019). The report describes RTE prospects in Estonia based on the definition, benefits, obstacles and risks of RTE. The definition of RTE in the report is as follows:

[1] In January 2020, as a researcher of the Cabinet Office of Japan, we visited Estonia to conduct a hearing survey for an overseas research study on the use of invoice information in the preparation of Supply and Use tables.

RTE is a digital ecosystem where transactions between diverse economic actors take place in or near real time by way of an increasingly automated exchange of digital, structured and machine-readable data in standardized formats. The resulting acceleration of information exchange and improved access to information is expected to reduce process latencies, save resources and transaction costs, increase organizational efficiency and business competitiveness, increase the speed and quality of decision-making, improve transparency, and stimulate economic and social innovation (Robert 2019).

In Estonia, the Estonian Association of Information Technology and Telecommunications (ITL)[2] lead the deployment of RTE under the name of real-time economy EE according to Vision 2030 (ITL 2018). The ITL will develop a system for real-time electronic invoicing and receipts transfer between organizations following European standards in 2018 as the first step in implementing RTE and plans to roll it out over the next ten years.

The government of Estonia divides RTE into three blocks: core technological infrastructure, E-service layer, and management layer, and describes the development stages of each block (Robert 2019). Of these, the core technological infrastructure and the E-service layer are currently the most developed areas, and there are many previous studies on various technologies and standard formats. On the other hand, the last one, the management layer, mainly focuses on the methodologies and concepts of accounting and auditing on a per-firm basis, and the discussion on a per-nation basis has not progressed much.

Let us look at the management layer concept; how to utilize obtained data in decision-making. Although the management layer's primary stakeholders are companies and the government, it mainly developed on the former because of the history of RTE. Companies' purpose for introducing RTE is more efficient resource management, better planning and risk assessment, and faster decision-making. On the other hand, the government's current vision is below.

"The same corporate data can also be used by state agencies to facilitate automated business reporting, real-time taxation or to compile national statistics without imposing reporting burdens on companies. Furthermore, as tools and technologies for data analytics and machine learning become increasingly prevalent, governments can make use of real-time data from diverse sources, such as national or third party databases or IoT ('internet of things') sensors, which would allow governments to build dashboards for continuously monitoring and assessing the country's economic situation and develop predictive models for forecasting economic events (e.g., company failures, changes in tax revenues) based on real-time data. This would allow governments to start providing customized services and feedback to companies (e.g. enabling companies to assess their indicators against their peers operating in the same sector or giving indications of possible risks) and to develop early warning systems for individual companies and the government (Robert 2019)".

[2] https://www.itl.ee/en/

However, when we visited Statistics Estonia to determine how many concrete measures they planned for implementing such management nationally, they said they were still conceptual and had not yet determined concrete methods. These concrete implementation methods will become an important research issue in the future and the production and operation of statistics. As an implementation method of this management layer, this paper proposes a method for dealing with accounting data for a nation and linking it to SNA, extending the concept of RTE, which deals with accounting data for a single company. Specifically, we propose a method for estimating Input–Output Tables (IOTs), one of the main statistics of SNA, based on invoice data, which is accounting data.

2 E-invoice data for statistics

This paper concerns the aggregation system and theories of the RTE's management layer that could be realized if detailed information, such as that contained in each invoice data, were made available in the SNA. The invoice system allows each company to record transaction information between sellers and buyers in the same data format. Theoretically, aggregating these transaction data provides information on supply chains among firms. SNA statistics, such as the IOT, aggregate the supply chain by industry and commodity. Thus, the transaction information between companies in the invoice could be the primary record of the flow concept in the SNA. Therefore, it is possible to calculate the flow of the intermediate demand part in the SNA by aggregating the transactions among all the companies included in the invoices. In addition, since each economic entity must submit tax-related data within a specified period, they are more preliminary and comprehensive than statistical surveys. Due to their characteristics, such as the speed of information updating and high completeness, there is growing interest in tax-related data worldwide, mainly for estimation purposes, such as quarterly preliminary GDP reports in European countries.

However, to realize this paper's claim and understand the supply chain, the statistical bureaus in each country must accumulate and use individual invoice data directly. Currently, no countries are using an estimation system based on individual invoices, and very few countries have established a system for collecting and using individual data for this purpose. Therefore, in the current situation, governments are generally expected to use not individual e-invoice data but aggregated Value Added Tax (VAT) total as a data source for quarterly GDP estimation.

Note that this paper is not a summary of the current systems in each country but a proposal of the systems and estimation theory necessary to realize the RTE in the future. This chapter summarises the current state of use as a preliminary step to the ideal set out in this paper. This discussion on the availability of individual invoice data and systems is dealt with in detail in the chapter Issues in the Statistical Use of Tax Information below.

EU countries and Australia already operate the invoice system and utilize the aggregated VAT information obtained from the invoice system for statistics. In Japan, the invoice system will start in October 2023, and each business operator will be obliged to keep invoices and other documents containing certain items to receive credit for purchase tax (NTAJ 2019). However, in Japan, the use of tax information for economic statistics is currently limited to the production of business statistics. There are various ways to use tax information, including invoices, for economic statistics. The Economic and Social Research Institute (ESRI), Cabinet Office of Japan, conducted an overseas research study in France (FR), Denmark (DK), Estonia (EE), and Australia (AU) in 2020 to investigate the use cases in other countries. The authors visited Estonia as a member of the study (ESRI 2023). Table 1 summarizes the use of VAT information in the surveyed countries for statistics, the Supply and Use Tables (SUT), the SNA tabulation procedure, and the linkage between statistical data and VAT data.

France and Denmark use VAT data to estimate investment and consumption in the SUT. In addition, France and Denmark compile individual VAT returns monthly to produce an economic index, the primary data for quarterly GDP estimation. France also includes VAT in the estimation of the use table in the annual estimation. Estonia and Australia do not directly use VAT data in the GDP estimation, but tax information is incorporated into the estimation as a reference value to identify trends in the quarterly estimation and for other statistical purposes. In Japan, all items are categorized as 'Not-utilized'.

2.1 Use in fraud detection

This paper focuses on the statistical use of invoice data. On the other hand, the most notable case of the utilization of invoice data is not economic statistics but the detection of VAT fraud. Especially in EU countries, economic losses related to VAT deductions due to Missing Trader Intra-Community (MTIC) fraud using non-existent operators and carousel fraud taking advantage of tax timing delays amount to several billions of euros per year (Lamensch 2018; Pouwels 2021). EU countries have taken measures such as introducing a VAT Information Exchange System (VIES) to share VAT information among member countries, but the losses have continued. Current fraud detection methods include Transactional Network Analysis (TNA), which analyzes networks and scores operators in the network based on VAT information. However, these methods have yet to reach a solution. As a result, TNA improves accuracy daily by updating and synchronizing data closer to real time.

The use of invoice data as a countermeasure against these frauds: digitization, data sharing, real-time data, and network analysis do not directly aim to utilize for economic statistics, which is the subject of this paper. However, these are a background for the rapid development of invoice data utilization and RTE infrastructure in the EU countries, simultaneously developing the infrastructure for utilization in statistics. As mentioned in the previous section, this paper aims to realize company-specific real-time monitoring, as claimed by the RTE, by using individual e-invoice data to understand the supply chain immediately. In economic statistical estimation, data for networked and immediate use of e-invoice data using such individual data needs to be made available.

Table 1 Current status of surveyed countries (Economic 2023)

	FR	DK	EE	AU
Data connection	Active	Active	Active	Partially
Statistics	Active	Active	Partially	Planning
SUT	Partially	Partially	Planning	Not-utilized

On the other hand, these are already in place and in use in the context of fraud detection, indicating a delay in statistics-related data linkage and system development.

3 e-Tax specifications and design

The difference between the data format of bookkeeping accounting data held by firms and the response format of questionnaires in statistical surveys increases the cost of responding to statistical surveys.

In Europe, on the other hand, the linkage between tax-related data and statistical surveys has been progressing. Europe established data standards and operational rules for e-tax in XML format as Pan-European Public Procurement Online (PEPPOL), and several countries provide an API for the reporting system. In Estonia, various accounting software is compatible with this API, and companies use the system in which necessary items are automatically estimated and reported from bookkeeping.

The specification of the API will enable accounting management companies to develop products that meet the specification and will automate administrative procedures related to statistical surveys and finance. In addition, if the government side provides an API that enables the automation of statistical surveys, the burden on respondents in responding to statistical surveys will be significantly reduced. At the same time, automatically and electronically compiled data will enable automatic estimation of statistics and real-time data collection.

Japan has also decided to join Open PEPPOL, the governing body of PEPPOL, to develop a standard specification for electronic invoicing. Currently, the Electronic Invoice Promotion Council is developing the BIS (Business Interoperability Specifications) specifications for the Japanese version of PEPPOL, and a draft of the specifications has been published (OpenPOPPOL 2021).

PEPPOL has established data standards in XML format and explains the advantages of these data standards as follows (OpenPOPPOL 2024):

The invoice and credit note provides simple support for complex invoicing, where there is a need for credit note in addition to an invoice. Other potential benefits are, among others:

- Can be mandated as a basis for national or regional eInvoicing initiatives.
- Procurement agencies can use them as basis for moving all invoices into electronic form. The flexibility of the specifications allows the buyers to automate processing of invoices gradually, based on different sets of identifiers or references, based on a cost/benefit approach.
- SME can offer their trading partners the option of exchanging standardised documents in a uniform way and thereby move all invoices/credit notes into electronic form.
- Large companies can implement these transactions as standardised documents for general operations and implement custom designed bi-lateral connections for large trading partners.
- Supports customers with need for more complex interactions.
- Can be used as basis for restructuring of in-house processes of invoices.

- Significant saving can be realised by the procuring agency by automating and streamlining in-house processing. The accounting can be automated significantly, approval processes simplified and streamlined, payment scheduled timely and auditing automated.

Readers can find detailed specifications on the official website. This paper describes the general specifications based on a example offerd by PEPPOL (OpenPOPPOL 2019).

XML (Extensible Markup Language) is a domain-specific language for structured documents. The basic structure of an XML document consists of elements, which are the building blocks of the document. Each element is enclosed in angle brackets and contains a name, which identifies the element, and optionally, attributes, which provide additional information about the element. The contents of an element can be text, other elements, or both. We could arrange elements in a tree structure, where parent elements contain child elements. This tree structure allows for the organization and representation of complex data in a standardized way.

The range between $< Invoice >$ and $< /Invoice >$ represents the information of one invoice. The example record two transactions, and the area between $< cac : IncoiceLine >$ and $< /cac : IncoiceLine >$ represents one transaction. Each transaction contains currency, product name, product description, item ID, item classification, a unit of measure, and the transaction price. Units and commodity classifications are defined based on the various classifications in each country. For instance, Japenese PEPPOL BIS uses Item type identification code (UNCL7143) as a item classification.

The entities of transactions are suppliers and customers, and the ranges enclosed by

$< cac : AccountingSupplierParty >... < /cac : AccountingSupplierParty >$

and

$< cac : AccountingCustomerParty >...< cac : AccountingCustomerParty >$

describe the information of each of them. The information of the subject of the transactionincludes name, company ID, country, address, telephone number, and e-mail address.

This paper assumes that economic bureaus could use directly e-invoices recorded in this XML data format for statistical estimation. As mentioned in previous chapters, the environment for the use of such data has not yet been achieved, and a number of issues need to be addressed before such data can actually be used in statistical estimation. Later, chapters discuss an analysis and survey of these challenges and the current situation.

4 Composition of the Input–Output Table

In this section, we try to create an IOT using invoice data. We tabulate the intermediate inputs by industry, according to the examples of calculations in official documents. However, when compiling the IOT, we must deal with various processes related to intra-enterprise transactions, margins, VAT, inventory, and price fluctuations. In addition, when we estimate some SNA concepts, we need to consider linkages between invoices and other data. The details of these considerations would exceed the volume limit of this paper even if we treated only a single concept. Therefore, in this paper, we only show the computability by calculating simple inter-industry transaction values rather than the concept of the intersection of intermediate demand in the actual SNA. In addition, this study assumes complete XML data and does not deal with missing data. In actual design, it is necessary to define alternative calculations and compensations for missing data, but this study presents examples of calculations based on hypothetical complete data.

4.1 State space for bookkeeping accounting

Although we usually calculate accounting information in a table format, it is unsuitable for concrete calculation procedure descriptions. This paper aims to convert the accounting information in XML format contained in invoice data into various formats, such as IOTs and economic analysis. Therefore, we need a description method suitable for the data calculation and estimation procedure, apart from the usual accounting table style description. By the way, programmers usually extract and compute data in XML format using a specific programming language, such as Java or Python, along with libraries. However, selecting a specific programming language and library for description language can compromise readability and generality for unfamiliar readers. Conversely, describing the estimation process of SNA using natural language and simple arithmetic, which is often adopted, can lead to a loss of accuracy and rigor in describing the complex process of transforming accounting information into SNA. Therefore, this paper adopts algebra as the descriptive language for the estimation process. While several mathematical methods describe accounting processes, this paper utilizes Exchange Algebra, an algebraic representation of accounting calculations (Deguchi 2004). Exchange Algebra retains the computational form of double-entry bookkeeping and allows us to perform bookkeeping calculations mathematically. Since Exchange Algebra consists of fundamental algebra, readers with a basic understanding of mathematics will find it straightforward to translate the procedures into specific algorithms or programs. The following sections will introduce, define, and explain the estimation process step by step.

We define a set of exchange algebra base elements as below:

$$
\begin{aligned}
&BaseElm = A \cup C \cup E \cup T \cup U \cup \{\#\}\\
&\quad where\\
&\quad A : \text{a set of Account titles}\\
&\quad C : \text{a set of Commmodity index}\\
&\quad E : \text{a set of Transaction Entity index}\\
&\quad T : \text{a set of Time period}\\
&\quad U = \{Yen, Amount\} : \text{a set of counting units}\\
&\quad \{\#\} : wildecard
\end{aligned}
$$

And also, we define a set of Exchange Algebra Base as below:

$$
\begin{aligned}
Ex\,Base = {} & A \cup \{\#\} \times C \cup \{\#\} \times E \cup \{\#\}\\
& \times T \cup \{\#\} \times U \cup \{\#\}
\end{aligned}
$$

Hereafter, an element of ExBase is denoted as $\langle a, c, e, t, u \rangle \in ExBase, a \in A, c \in C, e \in E$, $t \in T, u \in U$. Although this paper defines only the minimum required exchange algebra base elements, we can add arbitrary information such as region and VAT. In this case,

# is a wildcard and the following binary relation $R \subset Base\,Elm \times Base\,Elm$ is defined as below:

$$\begin{aligned}&(\forall x \in Base\,Elm, \#Rx \wedge xR\#)\wedge\\&\quad(\forall x, y \in S, \forall S \in \{A, C, E, T, U\}, xRy \vee yRx \Rightarrow x = y)\end{aligned}$$

Similarly, we introduce the binomial relation $R2 \subset Ex\,Base \times ExBase$ as below:

$$\begin{aligned}&\forall\langle a, c, e, t, u\rangle, \langle a', c', e', t', u'\rangle \in Ex\,Base,\\&\langle a, c, e, t, u\rangle R2\langle a', c', e', t', u'\rangle\\&\Rightarrow aRa', cRc', eRe', tRt', uRu'\end{aligned}$$

Hereafter, we will deal with the exchange algebra Ex with ExBase as the basis set.

We also introduce HatExBase and HEB as bellow:

$$\begin{aligned}HatExBase &= \{\hat{e}|e \in ExBase\}\\HEB &= ExBase \cup HatExBase\end{aligned}$$

We define a function $base : HEB \rightarrow ExBase$ that remove a hat operation as follows:

$$base[e \in HEB] = \begin{cases}\hat{}e & if\ e \in HatExBase\\ \ e & if\ e \in ExBase\end{cases}$$

Thus, we could introduce binary relation which represent equality on HEB defineded as follows:

$$\begin{aligned}&\forall e1, e2 \in HEB, e1\ R3\ e2\\&\quad\Rightarrow base[e1]\ R2\ base[e2]\\&\quad\wedge (e1, e2 \in HatEx\,Base \vee e1, e2 \in ExBase)\end{aligned}$$

Let ExElm denote a set of scalar multiple of $e \in HEB$, where $ExElm = \mathbb{R}^{+0} \times HEB$. Let be following function $eq : ExElm \times HEB \rightarrow HEB$ and $hatEq : ExElm \times HEB \rightarrow HEB$ that determines the equivalence relation on ExElm by R2 and returns 0 if there is no equivalence relation. Note that we denote 0 here by the product of $0 \times e \in ExElm.$

$$\begin{aligned}eq[xe][a] &= \begin{cases}xe & if\ base[e]\ R3\ base[a]\\ 0 & otherwise\end{cases}\\hatEq[xe][a] &= \begin{cases}xe & if\ e\ R3\ a\\ 0 & otherwise\end{cases}\\&where\\&xe \in ExElm\\&a \in HEB\end{aligned}$$

We define the function $norm : ExElm \rightarrow \mathbb{R}^{+0}$ to return the result of the norm calculation in the exchange algebra for scolar multiples of HEB as follows:

$$norm[a \times e] = a$$

Let Ex denote a set of linear combination of the element of ExElm, where

$$\begin{aligned}&x = a_1e_1 + b_1\hat{e}_1 + a_2e_2 + b_2\hat{e}_2 + ...a_ne_n + b_n\hat{e}_n \in Ex,\\&\quad a_1e_1, b_1\hat{e}_1, a_2e_2, b_2\hat{e}_2, ..., a_ne_n, b_n\hat{e}_n \in ExElm\end{aligned}$$

For the sake of simplicity we also introduce the following notation:

$$\begin{aligned}x &= a_1e_1 + b_1\hat{e}_1 + a_2e_2 + b_2\hat{e}_2 + ... + a_ne_n + b_n\hat{e}_n\\&= \langle\langle a_1, b_1\rangle, \langle a_2, b_2\rangle, ..., \langle a_n, b_n\rangle\rangle\end{aligned}$$

The Redundant algebra $Red(Ex, +, *, \hat{}, \bar{})$ over the Exchange Algebra (Ex) determined by the Hat ($\hat{}$) and Bar ($\bar{}$) operations is defined as follows:

$$\begin{aligned}&x = \langle\langle a_1, b_1\rangle, \langle a_2, b_2\rangle, ..., \langle a_n, b_n\rangle\rangle\\&\hat{x} = \langle\langle b_1, a_1\rangle, \langle b_2, a_2\rangle, ..., \langle b_n, a_n\rangle\rangle\\&\bar{x} = \langle\langle c_1, d_1\rangle, \langle c_2, d_2\rangle, ..., \langle c_n, d_n\rangle\rangle\\&\quad where\\&\quad c_i = a_i - b_i, d_i = 0 \; if \; a_i \geq b_i, i = 1, ..., n\\&\quad d_i = a_i - b_i, c_i = 0 \; if \; a_i < b_i, i = 1, ..., n\end{aligned}$$

$length[x] = n$ describe its length. Note that we denote $length[0] = 0$. We also define a function $xn : Ex \times \mathbb{Z}^{0+} \rightarrow ExElm$ to get the element by index as follows:

$$xn[x][n] = \begin{cases} 0 & if \; x = 0 \vee n = 0\\ 0 & if \; n > length[x]\\ a_ne_n & otherwise\end{cases}$$

We define a function $tail : Ex \rightarrow Ex$ to extract a subspace from $x \in Ex$ as follows:

$$tail[x] = \begin{cases} 0 \quad if \; length[x] \leq 1\\ \sum_{i=2}^{length[x]} xn[x][i] \quad otherwise\end{cases}$$

Then, we can define a projection function $projectEx : Ex \times HEB \rightarrow Ex$ and $projectHatEx : Ex \times EHB \rightarrow Ex$ as below:

$$\begin{aligned}&projectEx[x][y]\\&= \begin{cases} eq[xn[x][1]][y] \; if \; length[x] \leq 1\\ eq[xn[x][1]][y]\\ +projectEx[tail[x]][y] \; otherwise\end{cases}\\&projectHatEx[x][y]\\&= \begin{cases} hatEq[xn[x][1]][y] \; if \; length[x] \leq 1\\ hatEq[xn[x][1]][y]\\ +projectHatEx[tail[x]][y] \; otherwise\end{cases}\end{aligned}$$

Assuming that the transactions recorded in the invoice data over multiple years are $Flow \subset Ex$, the entire market transactions $flow[q] \in Flow$ in period q are obtained as follows:

$$flow[t] = \sum_{x \in Flow} projectEx[x][< \#, \#, \#, t, \# >]$$

4.2 Transformation from accounting state space to IOT

We constructed a state space using accounting information from invoices. Next, we will estimate an IOT based on the accounting information represented in Exchange Algebra. We describe the conversion from XML to exchange algebra for space limitation reasons in the Appendix.

For instance, consider a scenario where Company 1 and Company 2 exist. In period *t*, if Company 1 sells x pieces of goods a to Company 2 for x' yen, we denote this transaction as

$$\begin{aligned} y =& \hat{x}\langle Products, a, 1, t, Amount\rangle \\ &+x' \ \langle Cash, \#, 1, t, Yen\rangle \\ &+x \ \ \langle Products, a, 2, t, Amount\rangle \\ &+\hat{x'}\langle Cash, \#, 2, t, Yen\rangle \end{aligned}$$

Table 2 presents this description as double-entry bookkeeping, displayed in a table format.

Since the invoice data includes quantity and price valuations, we must define a conversion between them to preserve information redundancy. In transaction y above, although described by quantity valuation, extracting the conversion between price and quantity from the invoice data is straightforward. Individual transactions per product can detail the granularity of the price information in the invoice data. However, using average values by company, product, or period is also feasible, depending on the application. In this paper, we refrain from explicitly defining these units; instead, we utilize the 'toPrice' function. This function converts quantity evaluations to price evaluations based on the relationship between the commodity and its price. For the previously mentioned transaction y, if we suppose that the price of commodity a per unit is 2 Yen, we assume that "toPrice" converts the transaction as follows:

$$\begin{aligned} Price[y] =& 2\hat{x}\langle Product, a, 1, t, Yen\rangle \\ &+2x \ \langle Cash, \#, 1, t, Yen\rangle \\ &+2x \ \langle Product, a, 2, t, Yen\rangle \\ &+2\hat{x}\langle Cash, \#, 2, t, Yen\rangle \end{aligned}$$

Assume that in period q of a simplified market comprising four firms (1, 2, 3, 4), six commodities (a, b, c, d, e, f), and three industries (A, B, C), we observe the *flows*[*t*]:

Table 2 Transaction of x between Company 1 and Company 2

Company 1		
Term	Debit	Credit
t	a x	Cash x′
Company 2		
t	Cash x′	a x

$$
\begin{aligned}
flow[q] = {} & \hat{x}\langle Product, a, 1, t, Amount\rangle + x' \ \langle Cash, \#, 1, q, Yen\rangle \\
& + x \ \langle Product, a, 2, t, Amount\rangle + \hat{x'}\langle Cash, \#, 2, q, Yen\rangle \\
& + \hat{y}\langle Product, b, 1, t, Amount\rangle + y' \ \langle Cash, \#, 1, q, Yen\rangle \\
& + y \ \langle Product, b, 2, t, Amount\rangle + \hat{y'}\langle Cash, \#, 2, q, Yen\rangle \\
& + \hat{z}\langle Product, a, 1, t, Amount\rangle + z' \ \langle Cash, \#, 1, q, Yen\rangle \\
& + z \ \langle Product, a, 3, t, Amount\rangle + \hat{z'}\langle Cash, \#, 3, q, Yen\rangle \\
& + \hat{w}\langle Product, c, 3, t, Amount\rangle + w' \ \langle Cash, \#, 3, q, Yen\rangle \\
& + w \ \langle Product, c, 2, t, Amount\rangle + \hat{w'}\langle Cash, \#, 2, q, Yen\rangle \\
& + \hat{s}\langle Product, d, 3, t, Amount\rangle + s' \ \langle Cash, \#, 3, q, Yen\rangle \\
& + s \ \langle Product, d, 2, t, Amount\rangle + \hat{s'}\langle Cash, \#, 2, q, Yen\rangle \\
& + \hat{t}\langle Product, d, 3, t, Amount\rangle + t' \ \langle Cash, \#, 3, q, Yen\rangle \\
& + t \ \langle Product, d, 4, t, Amount\rangle + \hat{t'}\langle Cash, \#, 4, q, Yen\rangle
\end{aligned}
$$

Figure 1 depicts flow[q] as a network and illustrates the process of aggregating information from inter-firm transaction data to the IOT. We omit the descriptions of final demand and bookkeeping on the supplier side.

We obtain the purchases of a particular company for a particular commodity as follows. However, it is essential to note that this extraction method is valid only when Ex records no other accounts besides purchases and sales. If we consider other accounting treatments, defining a new Exchange Algebra Base and its element for these accounts becomes necessary:

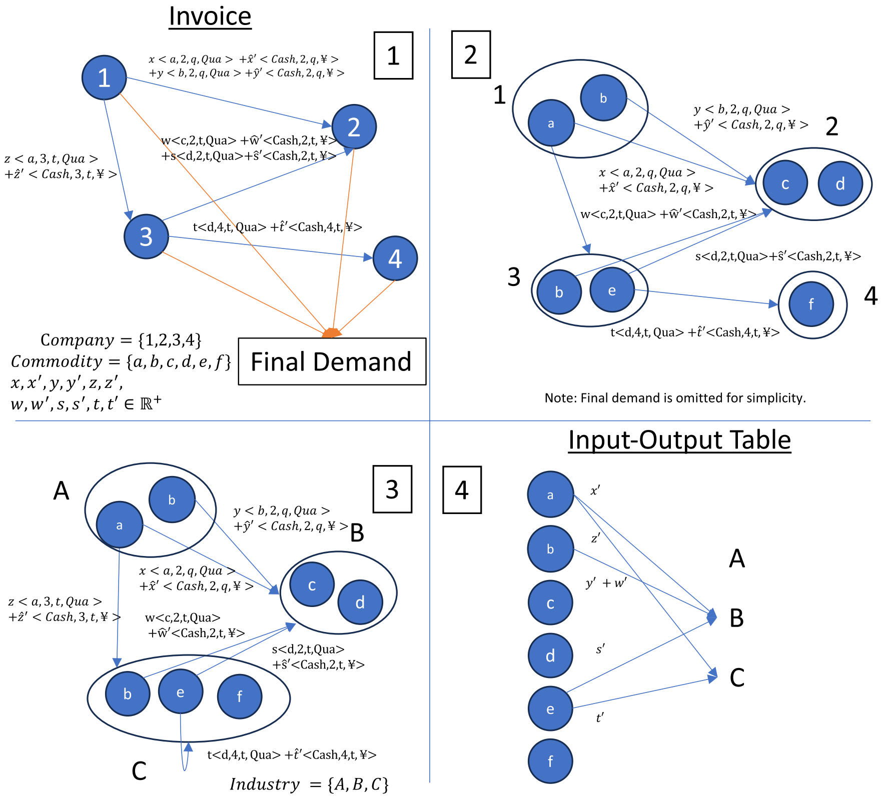
Invoice
1
1
2
3
4
$x < a, 2, q, Qua > +\hat{x}' < Cash, 2, q, ¥ >$
$+y < b, 2, q, Qua > +\hat{y}' < Cash, 2, q, ¥ >$
$z < a, 3, t, Qua >$
$+\hat{z}' < Cash, 3, t, ¥ >$
w<c,2,t,Qua> +ŵ′<Cash,2,t, ¥>
+s<d,2,t,Qua>+ŝ′<Cash,2,t, ¥>
t<d,4,t, Qua> +t̂′<Cash,4,t, ¥>
Final Demand
$Company = \{1,2,3,4\}$
$Commodity = \{a, b, c, d, e, f\}$
$x, x', y, y', z, z',$
$w, w', s, s', t, t' \in \mathbb{R}^+$
2
1
a
b
$y < b, 2, q, Qua >$
$+\hat{y}' < Cash, 2, q, ¥ >$
2
c
d
$x < a, 2, q, Qua >$
$+\hat{x}' < Cash, 2, q, ¥ >$
w<c,2,t,Qua> +ŵ′<Cash,2,t, ¥>
s<d,2,t,Qua>+ŝ′<Cash,2,t, ¥>
3
b
e
f
4
t<d,4,t, Qua> +t̂′<Cash,4,t, ¥>
Note: Final demand is omitted for simplicity.
3
A
a
b
$y < b, 2, q, Qua >$
$+\hat{y}' < Cash, 2, q, ¥ >$
B
c
d
$x < a, 2, q, Qua >$
$+\hat{x}' < Cash, 2, q, ¥ >$
$z < a, 3, t, Qua >$
$+\hat{z}' < Cash, 3, t, ¥ >$
w<c,2,t,Qua>
+ŵ′<Cash,2,t, ¥>
s<d,2,t,Qua>
+ŝ′<Cash,2,t, ¥>
b
e
f
C
t<d,4,t, Qua> +t̂′<Cash,4,t, ¥>
$Industry = \{A, B, C\}$
Input-Output Table
4
a
b
c
d
e
f
x'
z'
$y' + w'$
s'
t'
A
B
C

Fig. 1 Example of aggregation process from invoice to IOT

$$\begin{aligned}
&purchase[Ex][a \in C][e \in E][t \in T] = \\
&bar[toPrice[projectHatEx[Ex][X]]] \\
&where \\
&X =< Product, a, e, t, Amount >
\end{aligned}$$

For reasons of space, we denote $purchase[flow[t]][a][e][t]$ by $pf[a][e][t]$ in the following.

We calculate the weighted directed adjacency matrix of $Commodity \times Company$ network shown in Fig. 1 as follows:

$$\begin{bmatrix} pf[a][1][t] & \dots & pf[a][4][t] \\ \dots & \dots & \dots \\ pf[f][1][t] & \dots & pf[f][4][t] \end{bmatrix}$$

To estimate the $Commodity \times Industries$ IOT, we expand the economic entity E in ExBase to $E \cup Ind$, where Ind is a set of Industries. Therefore, we got a new ExBase, where

$$\begin{aligned}
Ex\,Base = {} & A \cup \{\#\} \\
& \times C \cup \{\#\} \\
& \times E \cup Ind \cup \{\#\} \\
& \times T \cup \{\#\} \\
& \times U \cup \{\#\}
\end{aligned}$$

We implement the transfer concept in accounting as a function $trans\,fer : Ex\,Base \times ExBase \times Ex \rightarrow Ex$:

$$\begin{aligned}
&transferExBase[b1][b2][b3] \\
&= \begin{cases} f[b2][b3], \ if \ b1R2\,b3 \\ b3, \ otherwise \end{cases} \\
&where, \\
&f[\langle a, c, e, t, u\rangle][\langle a', c', e', t', u'\rangle] \\
&= \langle g[a][a'], g[c][c'], g[e][e'], g[t][t'], g[u][u']\rangle \\
&g[w \in BaseElm][z \in BaseElm] \\
&= \begin{cases} \#, & if \ w = \# \wedge z = \# \\ w, & if \ w \neq \# \wedge z = \# \\ z, & if \ w = \# \wedge z \neq \# \\ z, & if \ w \neq \# \wedge z \neq \# \end{cases}
\end{aligned}$$

$$\begin{aligned}
&transferExElm[b1][b2][ab3] \\
&= \begin{cases} a\hat{}(transferExBase[base[b1]][b2][b3]) \ , \\ if \ b1 \in HatExBase \\ a(transferExBase[b1][b2][b3]), \ otherwise \end{cases} \\
&transfer[b1][b2][x] \\
&= \begin{cases} transferExElm[b1][b2][xn[x][1]], \\ \ if \ length[x] \leq 1 \\ transferExElm[b1][b2][xn[x][1]] \\ +transfer[b1][b2][tail[x]], \ otherwise \end{cases}
\end{aligned}$$

For simplicity, we denote
$transfer[b3][b4][transfer[b1][b2][x]]$ by
$transfer[\langle b1, b3\rangle][\langle b2, b4\rangle][x]$.
Then, we obtained a $Commodity \times Industrynetwork$ as below:

$$\begin{aligned}
&indFlow[t] = transfer[B][C][flow[t]]\\
&= x\hat{}\langle Product, a, A, t, Amount\rangle + x'\ \langle Cash, \#, A, t, Yen\rangle\\
&+ x\ \langle Product, a, B, t, Amount\rangle + x'\hat{}\langle Cash, \#, B, t, Yen\rangle\\
&+ y\hat{}\langle Product, b, A, t, Amount\rangle + y'\ \langle Cash, \#, A, t, Yen\rangle\\
&+ y\ \langle Product, b, B, t, Amount\rangle + y'\hat{}\langle Cash, \#, B, t, Yen\rangle\\
&+ z\hat{}\langle Product, a, A, t, Amount\rangle + z'\ \langle Cash, \#, A, t, Yen\rangle\\
&+ z\ \langle Product, a, C, t, Amount\rangle + z'\hat{}\langle Cash, \#, C, t, Yen\rangle\\
&+ w\hat{}\langle Product, c, C, t, Amount\rangle + w'\ \langle Cash, \#, C, t, Yen\rangle\\
&+ w\ \langle Product, c, B, t, Amount\rangle + w'\hat{}\langle Cash, \#, B, t, Yen\rangle\\
&+ s\hat{}\langle Product, d, C, t, Amount\rangle + s'\ \langle Cash, \#, C, t, Yen\rangle\\
&+ s\ \langle Product, d, B, t, Amount\rangle + s'\hat{}\langle Cash, \#, B, t, Yen\rangle\\
&+ t\hat{}\langle Product, d, C, t, Amount\rangle + t'\ \langle Cash, \#, C, t, Yen\rangle\\
&+ t\ \langle Product, d, C, t, Amount\rangle + t'\hat{}\langle Cash, \#, C, t, Yen\rangle\\
&where,\\
&B = \langle\langle \#, \#, 1, \#, \#\rangle, \langle \#, \#, 2, \#, \#\rangle\\
&\qquad , \langle \#, \#, 3, \#, \#\rangle, \langle \#, \#, 4, \#, \#\rangle\rangle\\
&C = \langle\langle \#, \#, A, \#, \#\rangle, \langle \#, \#, B, \#, \#\rangle\\
&\qquad , \langle \#, \#, C, \#, \#\rangle, \langle \#, \#, C, \#, \#\rangle\rangle
\end{aligned}$$

For reasons of space, we denote $purchase[indFlow[t]][a][e][t]$ by $pi[a][e][t]$ in the following. We calculate the weighted directed adjacency matrix of $Commodity \times Industory$ network shown in Fig. 1 as follows:

$$\begin{bmatrix} pi[a][A][t] & ... & pi[a][C][t] \\ ... & ... & ... \\ pi[f][A][t] & ... & pi[f][C][t] \end{bmatrix}$$

Therefore, we can obtain the IOT by calculating the norm from each purchase:

$$\begin{bmatrix} norm[pi[a][A][t]] & ... & norm[pi[a][C][t]] \\ ... & ... & ... \\ norm[pi[f][A][t]] & ... & norm[pi[f][C][t]] \end{bmatrix}$$

4.3 Extension to company-based monitoring

So far, we estimate the IOT by transforming the state space of each entity's accounting records into a simple scalar quantity using the norm operation. This transformation means we have removed most of the accounting information in the Exchange Algebra, reducing it to simple quantities.

Wassily Leontief established the IOT as a commodity-industry matrix in the 1930 s. This structure resulted from technical constraints such as data availability and computational resources rather than theoretical limitations. To facilitate the application of the theory under these constraints, he introduced the industrial technology assumption and the commodity technology assumption.

The industrial technology assumption posits that identical industries have identical production functions, implying infinite substitutability of materials within the same industry. This leads to neglecting concepts such as distribution costs, diversity, and competition among firms in the analysis. Similarly, the commodity technology assumption assumes no regional product differences, such as price, transportation, and time costs. In addition, invoices contain information on the actual time of each economic transaction. However, the IOT aggregates information on individual transactions and expresses them as inputs and outputs over a certain period, lacking the concept of the order or timing of transactions.

Since invoices contain geographical and temporal information on transactions, using them directly allows us to observe accounting information changes by region and firm over time. Similarly, by considering invoices as book records, we can go beyond the fundamental input–output relationship and observe information updates in accounting event units such as profit, loss, inventory, capital accumulation, and depreciation.

Therefore, originally we should not apply the norm operation and we should treat all the information contained in the invoice as a state space as it is. This is precisely the kind of data RTE aims for, allowing economic analysis on a company basis based on accounting information.

Let $state[t][e] \in Ex$ denote state space of entity e at end of period t, where

$$\begin{aligned} state[t][e] = {} & bar[state[t-1][e] \\ & + projectEx[flow[t]][\langle \#, \#, e, t\rangle]] \end{aligned}$$

With $flow[t]$ above, we can obtain $state[t][e]$ as follows.

$$\begin{aligned} state[t][1] = {} & \sum_{0 \leq s \leq t-1} state[s][1] \\ & + \hat{x}\langle Product, a, 1, t, Amount\rangle \\ & + x' \ \langle Cash, \#, 1, t, Yen\rangle \\ & + \hat{y}\langle Product, b, 1, t, Amount\rangle \\ & + y' \ \langle Cash, \#, 1, t, Yen\rangle \end{aligned}$$

Various forms of expression can be used to realize the RTE's goal of monitoring changes in financial information by company, but no methodology can be used other than constructing a state space related to accounting in this way and extracting and aggregating the appropriate information.

By recording and updating information in this manner, we can ascertain the changes in the state of each economic entity due to economic transactions. However, although this model could capture only the results of the transactions recorded on the invoice, changes in accounting information associated with various economic activities, such as production and interest payments, cannot be comprehended. Although it is possible to construct a simulation model that includes such information, we need to extend the model to account for the previous period's state of the economy, production constraints such as labor and capital, and external variables such as demand, prices, and interest rates. In addition, to address the economy's state, we must incorporate accounting treatments such as wage payments, tax collection, inventory valuation adjustments, depreciation, and production needs into the model. This extension, significantly different from the existing SNA, is beyond the scope of this paper and will not be addressed here; it will be a subject for future research.

5 Issues in the statistical use of invoice data

Implementing the electronic invoice system is an excellent opportunity to obtain new data for statistics, but we need to solve several problems for the statistical use of it. This paper employs Japanese invoices and yen as the currency unit for convenience. However, this methodology is not exclusive to Japan; it is particularly relevant to European countries with established electronic invoicing systems by PEPPOL. We could divide problems into technical and political ones. Although both are hard, the political issues will be challenging to realize a certain technical ideal previously described.

5.1 Data storage format

First, the essential issue is the discussion on the data storage format. The Japanese government has made it mandatory for businesses to store electronic invoices. However, documents such as PDFs are allowed as a storage format and are not limited to machine-readable formats.

The Estonian government has not mandated the issuance of B2B e-invoices, and as of 2019, only 72% of companies in Estonia use or partially use e-invoices. However, to realize RTE, they set the electronic and automated processing of all transactions between business parties as a goal.

Therefore, since 2017, the introduction of machine-readable accounting documents and submitting annual financial reports in XBRL format have been promoted. In addition, the amendment of the accounting law introduces the obligation for all suppliers to public authorities to issue e-invoices from July 1, 2019. Finland, which is also committed to RTE, started using e-invoice in 2010, and its use in the private sector is more widespread than in Estonia. The Danish government has mandated using PEPPOL for all B2G transactions since 2022 and is considering making it mandatory for B2B.

One of the essential advantages of RTE is the increased efficiency due to automation. In an interview with the Estonian Institute of Accountants, they predict that the introduction and development of electronic invoicing will enable companies to automate a large part of their accounting processes, turning accounting operations, which currently account for less than 1% of Estonia's GDP, into a more efficient value-adding business (Robert and Tarmo 2019). Similarly, the project also looks at automating payroll, tax-related processing, and accounting reports between the government, banks, and businesses. However, these are only the benefits of introducing electronic invoicing, and we could not receive the benefits if we allow using paper or PDF storage formats, as is the case in Japan.

It will be a significant loss in future e-government and digital data utilization if we do not discuss the secondary use and overall data integration of newly available electronic data, such as electronic invoices.

5.2 Legislation and institutional systems

Even in countries that introduced invoice systems, there are challenges in utilizing invoice data for statistical purposes. These challenges include the legal system permitting statistical use, the frequency of data submission, and the obligation to submit to the government. For instance, the current Japanese legal system does not permit invoice data for statistical estimation, as this constitutes use for purposes other than those intended. In addition, while there is an obligation to retain invoices, there is no system to ensure that these invoices are submitted to the tax authorities regularly.

This paper presents the theoretical possibilities for using invoice data in SNA. It does not propose the immediate realization of a RTE in Japan or any particular country. Instead, the following sections will present examples of countries that have already established the institutional foundations for linking invoice data with statistics, suggesting that the recommendations in this paper are not pipe dreams but rather practical steps towards a more efficient economic system.

An example of a country where institutional issues have been resolved is the Republic of Korea. Although the Republic of Korea has not introduced Peppole, since 2015, the National Tax Service of Korea has been providing the Hometax system, which integrates internet services related to taxation into a single platform (OECD 2023). In addition, the Tax Statistics DataBase has been established by collecting purchase tax statements from business establishments. With this system, enterprise statistics are already produced in Korea using tax data from all enterprises, ensuring comprehensive and high coverage in statistical analysis, thereby providing reliable and robust data for economic research (Eurostat 2021).

Furthermore, tax authorities verify or approve data before the recipient receives an electronic invoice in countries that have implemented the Peppol CTC (Continuous Transaction Controls) Method (EDICOM 2024). The government collects and stores the data concurrently with the issuance of the invoice. These systems are widely used in Latin American countries and have been introduced in Saudi Arabia and Israel. The deadlines for submitting and updating data vary, such as monthly or annually, depending on the country (e.g., SAF-T in Poland, Norway, and Lithuania; E-Accounting in Portugal). An example of CTC with real-time reporting is the RTIR in Hungary.

Since these countries have already resolved the institutional issues related to the statistical use of invoice data, they can realize a RTE by applying the theoretical estimation methods presented in this paper to the data formats in each country.

However, even in countries with well-developed e-government systems, establishing the legal basis and conditions for public organizations to access private business data for statistical production and analysis is challenging. In addition, technical verification of what kind of information computation is compatible with protecting anonymity and establishing regulations are also significant issues.

Furthermore, before establishing a legal system, it is necessary to address the public's distrust of tax and customs authorities and statistical offices, their wariness of increased state control, and their concern about the increased workload associated with reporting. In particular, the technical advantages of using statistics are more specialized than reporting automation and may not be easily understood.

5.3 Coordination of data linkages and concepts

As we showed in the previous chapter, it is theoretically possible to construct SNA from invoices and accounting information. However, in the actual estimation, it is necessary to ensure the consistency of the concept with existing statistics. Furthermore, it is necessary to sort out what information in the existing statistical concepts can be obtained by newly obtained data such as invoice, financial, and taxation data in the future.

5.3.1 Limitation of invoice data

Even if we limit our discussion to the IOT, the current invoice data cannot capture margins, transactions among businesses within the same company, private consumption, transactions with tax-exempt businesses, final demand, and value-added. For these ungraspable items, it is necessary to design a new method of use, such as using other big data like POS, or using the data to supplement the current statistical surveys.

For instance, regarding the issue of final demand, we can estimate it as a shortage of output of some nodes using total output and network information, demonstrated in another study (Ohosato et al. 2018). However, combining this approach with PoS (Point of Sales) data and existing consumption surveys would be more advantageous.

We must also address the commodity flow method, intermediate input goods, and capital goods classification. Our previous studies utilizing private-sector data resolved this issue by preemptively categorizing goods into intermediate consumption goods and capital goods (Akagi et al. 2015). Theoretically, we can classify goods regularly purchased as intermediate input goods with access to time series data. Although relying on such estimated calculations should generally be avoided in official statistics, given that the input surveys currently conducted by various countries are essentially based on sample surveys, the discussion of precision differences becomes relevant.

Since discussing the accuracy of these approaches without actual data is challenging, we need to discuss these topics with actual data in future work.

5.3.2 Data format for statistical use

In addition, like designing the survey format, it is necessary to devise the data format and so on based on the assumption that we will utilize data for statistical estimation. For example, the designers of invoice data select item information by focusing only on accounting categories, such as whether the item is subject to the reduced tax rate. Therefore, it is necessary to classify those transaction items into the existing item classifications, such as in the IOT. Much work would be required to sort out the correspondence between the industrial and commodity classifications and the UNCL7143 used in the invoices.

For actual use, it is necessary to design the item rating in the invoice according to the purpose, for example, to provide a format that allows selection by primary classification or to provide an input assistance function on the web. The various concepts in invoices, not limited to item classification, are inconsistent with the concepts in SNA, and there needs to be a discussion on what information is necessary for conversion. Currently, there is no concept of general data use and no coordination, and we need to establish such coordination and design for data users as soon as possible.

5.3.3 Area and establishments

One of the most significant merits of invoice data for statistical use is the low granularity of area information. If we use the regional information of the entities that issue invoices, we could obtain the transactions by region in units of cities, wards, towns, and villages, and we can estimate regional and inter-regional IOTs. If an invoice is issued, we can perform the same calculation for imports and exports to and from foreign countries. However, there is no guarantee that the entity issuing the invoice is the same as the production factory. The primary reason the SNA disaggregates by establishment rather than by enterprise is the difficulty of disaggregating by industry/commodity because of the wide range of activities undertaken by enterprises. However, if the industrial technology assumption were to be adopted, the input–output ratio of the data for enterprises that deal with a single product could be used for prorating.

Moreover, as demonstrated in our previous study, estimation of the IOT using Teikoku Databank data, enhancing establishment-based estimations by incorporating methods that consider employee numbers and distances between firms, is effective (Ohosato et al. 2018). This paper advocates merging such approaches with data sources like economic censuses and business registries instead of relying solely on single-data estimates from invoices. Therefore, addressing estimation-related issues becomes more straightforward when utilizing readily available census data, including employee numbers per establishment and sales details.

6 Conclusion

This paper has examined the methodology for utilizing invoice data in SNA and achieving a RTE. It acknowledges that governments globally have established systems for the electronic collection and processing of invoice data, which are increasingly being used for various statistical purposes.

Our findings indicate that invoice data can at least estimate the IOT in SNA and subdivide it into finer granularity regarding region, industry, product category, and time. Furthermore, we formulated a method for monitoring the financial situation of individual companies, one of the objectives of the RTE.

However, the institutional and technical environments pose significant obstacles to realizing the full potential of these invoice data.

The introduction of new data has the potential to fundamentally transform economic statistics. However, to effect such a change, we must address numerous challenges.

Political and institutional issues, in particular, cannot be resolved by researchers alone. Although researchers typically do not engage directly in political activities, they must propose theoretically viable options and visions. The concept of integrating big data into economic statistics remains ambiguous. This paper offers a vision for how new data could extend the capabilities of SNA, and we invite all stakeholders to join us in this transformative journey.

To implement the recommendations provided in this paper practically, it is essential to present, formulate, and implement new methods and the benefits they bring in a feasible manner. Achieving this involves overcoming several technical challenges. We are committed to focusing our future efforts on mapping concepts between data and SNA, developing prototypes, and designing systems for automation based on real-world data.

Appendix

Invoice data conversion

In this section, we define a transformation from invoice data to exchange algebra. The official document of PEPPOL contains only examples of calculations within a single invoice, and there is no notation for aggregating multipcle invoices.Thus, we introduce the following notation. Let EN be the set of Element Names in the XML of an invoice. If the invoice contains only the following description, the EN will show in the example below. Note that we omitted the end tag in XML in the EN.

$$\begin{array}{l}\{< cac : Country > \\ \quad < cbc : IdentificationCode >\} \\ \mathrm{GB} \\ \quad < cbc : IdentificationCode > \\ \quad < ac : Country >\end{array}$$

$$\begin{aligned}EN =&\{< cac : Country > \\ &, < cbc : IdentificationCode >\}\end{aligned}$$

Let D denote the set of data in XML. Although invoice data defines four types (Decimal, String, Date, and Binary), we consider them all as single abstract data.

In the previous example, D will be the following:

$D = \{GB\}$

Let DM be the map of D onto EN.

$DM \subset EN \times D$

In the previous example, DM is the following:

$$\begin{aligned}DM = &\{\langle < cbc : IdentificationCode > \\ &, GB >\rangle\}\end{aligned}$$

The tree structure consisting of EN is defined as follows:

$ENTree = \langle EN, E \rangle$

Note that EN is a vertex set, and E is a pair of edge sets.

$E \subset EN \times EN$

Using the previous example, we can obtain E as follows:

$$\begin{aligned} E = &\{\langle < cac : Country > \\ &, < cbc : IdentificationCode > \rangle\} \end{aligned}$$

In this case, we can obtain the terminal node of the tree structure as follows:

$$\begin{aligned} leaf[\langle EN, E \rangle] = &\{e2 | \langle e1, e2 \rangle \in E \\ &, (\forall x) \langle x, y \rangle \in E \\ &, x \neq e2\} \end{aligned}$$

The map between XML leaf and Data is defined as follows:

$$\begin{aligned} LDMap[\langle EN, E \rangle] = &\{\langle l, c \rangle | \langle l, c \rangle \in DM \\ &, l \in leaf[\langle EN, E \rangle]\} \end{aligned}$$

The structure consisting of ENTree and DM is called Elements in XML.

$Element = \langle EN, E, DM \rangle$

We define the set of Elements as XML. Elements with the same Tree structure and data are considered identical and therefore do not have a duplicate in XML. Also, subtrees within an Element do not have duplicate Element Names.

Let us consider the extraction of a subtree from an Element.

In this case, we define the following function to extract the Edge from a specific node that leads to its children:

$$\begin{aligned} &child[X \subset EN][\langle EN, E, DM \rangle] \\ &= \{\langle e, c \rangle | \langle e, c \rangle \in E, e \in X\} \end{aligned}$$

Functions to extract descendants are similarly defined as follows:

$$\begin{aligned} &descendant[X \subset EN][EED] \\ &= \begin{cases} \phi, & if\ X = \phi \\ A \cup B, & otherwise \end{cases} \\ &where, \\ &EED = \langle EN, E, DM \rangle \\ &A = child[X][EED] \\ &B = \\ &\{descendant[\{e\}][\langle EN, child[X][EED], DM \rangle] \\ &| \langle y, e \rangle \in child[X][EED]\} \end{aligned}$$

Thus, we can obtain a subtree of Element headed by a particular $e \in EN$ by bellow function:

$$
\begin{aligned}
& subTree[e \in EN][EED] \\
& = \langle EN \\
& , descendant[\{e\}][EED] \\
& , \{\langle l, d\rangle | \langle l, d\rangle \in DM \\
& , \exists\langle x, l\rangle \in descendant[\{e\}][EED]\}\rangle
\end{aligned}
$$

Hereafter, we denote $subTree[e2][subTree[e1][X]]$ by $subTree[\langle e1, e2\rangle][X]$ for simplicity. We consider a function to extract a particular element or subtree from XML.

$$
\begin{aligned}
& projectElm[e \in EN][XML] \\
& = \{subTree[e][x] | x \in XML\}
\end{aligned}
$$

Assuming the entire example of XML as SAMPLE, the calculation results are as follows:

$$
\begin{aligned}
projectData[\langle & < cac : InvoiceLine > \\
& , < cac : Price > \\
& , < cbc : PriceAmount > \rangle][SAMPLE] \\
& = \{400, 500\}
\end{aligned}
$$

Define a function to extract only specific data from XML as follows:

$$
\begin{aligned}
& projectData[\langle e_1, e_2, ..., e_n\rangle \in EN^n][XML] \\
& = \{d | \langle x, d\rangle \in DM, \langle X, E, DM\rangle \in Y\} \\
& where, \\
& Y = projectElm[\langle e_1, e_2, , ..., e_n\rangle][XML]
\end{aligned}
$$

With SAMPLE, it follows that,

$$
\begin{aligned}
projectData[\langle & < cac : InvoiceLine > \\
& , < cac : Price > \\
& , < cbc : PriceAmount > \rangle][SAMPLE] \\
& = \{400, 500\}
\end{aligned}
$$

Below, we define functions to extract elements of the exchange algebra basis from the invoice data to be converted to the exchange algebra using these functions. The function to extract a company ID from XML is defined as follows:

$$
\begin{aligned}
& supplierID[XML] = projectData[A][XML] \\
& where, \\
& A = \langle < Invoice > \\
& , < cac : AccountingSupplierParty > \\
& , < cac : Party > \\
& , < cac : PartyLegalEntity > \\
& , < cbc : CompanyID > \rangle
\end{aligned}
$$

$$\begin{aligned}
&customerID[XML] = projectData[A][XML]\\
&where,\\
&A = \langle < Invoice >\\
&, < cac : AccountingCustomerParty >\\
&, < cac : Party >\\
&, < cac : PartyLegalEntity >\\
&, < cbc : CompanyID > \rangle
\end{aligned}$$

The function to extract the transaction period is defined as follows. Although it is possible to use arbitrary aggregations such as quarter, month, week, and day, depending on the application, the function to extract the year from the YYYY-MM-DD format is defined as "year" and is used.

$$\begin{aligned}
&period[XML] = year[dueDate[XML]]\\
&where,\\
&dueDate[XML] = projectData[A][XML]\\
&A = \langle < Invoice >, < cbc : DueDate > \rangle
\end{aligned}$$

The function to extract an invoice line from XML is defined as follows: IL representing an invoice line is a subset of XML.

$$\begin{aligned}
&invoiceLines[XML] =\\
&projectElm[\langle < cac : InvoiceLine > \rangle][XML]
\end{aligned}$$

We define the following functions to extract transaction volume, monetary valuation, and commodity ID from IL.

$$\begin{aligned}
&quantity[XML] = projectData[A][IL]\\
&where,\\
&A = \langle < Invoice >< cbc : InvoicedQuantity > \rangle\\
&yen[IL] = projectData[A][IL]\\
&where,\\
&A = \langle < Invoice >\\
&, < cbc : LineExtensionAmount > \rangle\\
&commodityID[IL] = projectData[A][IL]\\
&where,\\
&A = \langle < Invoice >\\
&, < cac : Item >\\
&, < cac : StandardItemIdentification >\\
&, < cbc : ID > \rangle
\end{aligned}$$

From the above, the following functions define the conversion of XML to exchange algebra. These procedures convert invoices to double-entry bookkeeping, but note that VAT and other accounting procedures are omitted here for simplicity.

$$
\begin{aligned}
& xmlEx[XML] = \sum_{IL \in invoiceLine[XML]} x \\
& where, \\
& x = q\,\hat{\langle} Products, a, i, t, Amount\rangle \\
& \quad + y\langle Cash, \#, i, t, Yen\rangle \\
& \quad + q\langle Products, a, j, t, Amount\rangle \\
& \quad + y\,\hat{\langle} Cash, \#, i, t, Yen\rangle \\
& i = supplierID[XML] \\
& j = customerID[XML] \\
& p = period[XML] \\
& y = yen[IL] \\
& a = commodityID[IL] \\
& q = quantity[IL]
\end{aligned}
$$

$$
xlmsEx[XMLS] = \sum_{XML \in XLMS} xmlEx[XML]
$$

Acknowledgements

In writing this paper, we are grateful to Professor Motohiko Nakamura of the Chiba University of Commerce for providing us with the documents about the invoice system and its specifications. I am also grateful to Professor Hiroshi Deguchi and Professor Takao Terano for their advice on the paper's overall structure. I want to take this opportunity to express my gratitude to them. All errors in this paper are the responsibility of the author.

Author contributions

Not applicable.This paper was written with the contribution of one author.

Funding

The Article Processing Charge was covered by the funds of PAPAIOS and JSPS (KAKENHI Grant No. JP21HP2002). The Chiba University of Commerce Individual Research Fund financed this study.

Availability of data and materials

All data generated or analysed during this study are included in this published article.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 18 August 2023 Accepted: 27 August 2024

Published online: 15 October 2024

References

Akagi K, Ohsato T, Deguchi H (2015) Input-output table constructed with private business data and its algebraic description

Bernal I, Sejersen T (2021) Big data for economic statistics. Stats Brief 28, United Nations Economic and Social Commission for Asia and the Pacific. https://www.unescap.org/sites/default/d8files/knowledge-products/Stats_Brief_Issue28_Big_data_for_economic_statistics_Mar2021.pdf

Deguchi H (2004) Economics as an agent-based complex system: toward agent-based social systems sciences. Springer, Cham

ESRI (2023) Overseas survey research on the use of invoice information in the preparation of supply and use tables (Summary). https://www.esri.cao.go.jp/jp/esri/about/gdpstat_kaizen/sut_gaiyou.pdf

Eurostat: Korea-eu - business demography. Statistical report, European Commission (2021). https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Archive:Korea-EU_-_business_demography

EDICOM (2024) Electronic Invoicing Models: CTC, Clearance, Real-Time, Centralized, Interoperable, and More... Accessed: 2024-07-24. https://edicomgroup.com/blog/electronic-invoice-models-ctc-clearance-interoperability

ITL (2018) Vision of the Estonian Association of Information Technology and Telecommunications (ITL) of Information Society in 2030: SMART ESTONIA. https://www.itl.ee/en/vision-2030/

Itsuo S, Deguchi H, Omori A (2021) Alternative approaches to the axiomatisation of national accounting: As a tribute to the two great norwegian figures in the world of national accounting. the 36th IARIW Virtual General Conference

Kogan A, Sudit EF, Vasarhelyi MA (1990) Continuous online auditing: A program of research. J Inf Syst 13(2):87–103
Lamensch M (2018) Vat fraud: economic impact, challenges and policy issues. European Parliament, Strasbourg
Laney D (2001) 3D data management: controlling data volume, velocity, and variety. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
Mehta M, Julaiti J, Griffin P, Kumari S (2021) Big data research in fighting COVID-19: contributions and techniques. Big Data Cogn Comput 5(3):30. https://doi.org/10.3390/bdcc5030030
OECD: corporate tax statistics database (2023). https://www.oecd.org/tax/beps/corporate-tax-statistics-database.htm
Ohosato T, Akagi K, Deguchi H (2018) construction of an input-output table that takes into account business-to-consumer transactions using private data
Ohosato T, Akagi K, Deguchi H (2018) Developing an input-output table generation algorithm from a large scale company database in japan: How to deal with ambiguous export and import information
OpenPEPPOL (2021): Japanese PEPPOL BIS Documentation/Japan Standard Commercial Invoice process(BIS). https://test-docs.peppol.eu/poacc/billing-japan/
OpenPEPPOL (2024): Peppol BIS Billing 3.0/Business Interoperability Specifications(BIS). https://docs.peppol.eu/poacc/billing/3.0/bis/
OpenPEPPOL (2019) O.: OpenPEPPOL/peppol-bis-invoice-3/rules/examples/base-example.xml. https://github.com/OpenPEPPOL/peppo-bis-invoice-3/blob/master/rules/examples/base-example.xml
Pouwels A (2021) Missing trader intra-community fraud. European Parliament, Strasbourg
Robert K, Tarmo K (2019) Real-time economy: definitions and implementation opportunities
Robert K, Tarmo K, Art A, Maarja T, Ralf-Martin S, Carsten S (2019) Real-time economy: definitions and implementation opportunities, 68
Siegele L (2024) The real-time economy: how about now? https://www.economist.com/special-report/2002/02/02/how-about-now
NTAJ (2019) Understanding the reduced VAT rate system (In Japanese). https://www.nta.go.jp/taxes/shiraberu/zeimokubetsu/shohi/keigenzeiritsu/pdf/0018006-112.pdf
Vasarhelyi MA, Teeter RA, Krahel J (2023) Audit education and the real-time economy **25**(3), 405–423. https://doi.org/10.2308/iace.2010.25.3.405. Accessed 2021-01-07
Victoria C, Qi L, Vasarhelyi AM (2014) The development and intellectual structure of continuous auditing research **33**(1), 37–57
RARC (2023) The Real-time Economy: The Technological Basis for Reengineered Business Reporting. http://raw.rutgers.edu/node/29.html

Publisher's Note