# EURONA

**EU**ROSTAT **R**EVIEW
**O**N **N**ATIONAL **A**CCOUNTS
AND MACROECONOMIC
INDICATORS

2/2019

eurostat

# EUROSTAT REVIEW ON NATIONAL ACCOUNTS AND MACROECONOMIC INDICATORS

2/2019

*Printed by Imprimerie Bietlot in Belgium*
Manuscript completed in December 2019



Luxembourg: Publications Office of the European Union, 2020



Theme: Economy and finance
Collection: Statistical books

| | | |
|---|---|---|
| Print | ISSN 2443-7832 | Cat.: KS-GP-19-002-EN-C |
| PDF | ISSN 1977-978X | Cat.: KS-GP-19-002-EN-N |

# Contents

# Aims and scope

EURONA is an open access, peer-reviewed, scholarly journal dedicated to National Accounts and Macroeconomic Indicators. EURONA aims at providing a platform for researchers, scholars, producers and users of macroeconomic statistics to exchange their research findings, thereby facilitating and promoting the advancement of National Accounts and Macroeconomic Indicators.

EURONA publishes empirical and theoretical articles within the scope of National Accounts and Macroeconomic Indicators, as well as articles on important policy uses of these statistics. They may relate to both users' and producers' interests, present subjects of general relevance or investigate specific topics.

EURONA is non-partisan and applies the highest standards to its content, by emphasising research integrity, high ethical standards, validity of the findings and cutting edge results. EURONA gives room to all viewpoints.

The articles published in EURONA do not necessarily reflect the views or policies of the European Commission.

Website: https://ec.europa.eu/eurostat/web/national-accounts/publications/eurona

Contact: ESTAT-EURONA@ec.europa.eu

# Editors



# Editorial board

# Editorial

In 2005, the late Sir Tony Atkinson produced, on the invitation of the United Kingdom's National Statistician, a report with recommendations on the measurement of government output and productivity in the national accounts. Since that time, a lot of research has been undertaken and new — output-based — methods have been introduced in many countries. The report also had an impact on the guidance given in the SNA 2008 on the measurement of non-market output. In the first article of this issue of EURONA, Fred Foxton, Joe Grice, Richard Heys and James Lewis review the work carried out and the results obtained in the United Kingdom following the implementation of the approach proposed by Sir Tony Atkinson.

The second paper is globalisation-related: Bernhard Michel, Caroline Hambÿe, Bart Hertveldt and Guy Trachez present the progress made in Belgium on the development of extended supply and use tables. These tables provide deeper breakdowns of industries by aspects related to globalisation, such as domestic or foreign ownership of enterprises, share of output that is exported and firm size. These tables thus provide a better framework for the analysis of global value chains. The paper also discusses the various choices to be made and challenges to be overcome in the further development of extended supply and use tables, which will be very useful for other countries that aim to compile similar data.

The third and fourth papers both relate to the use of new data sources in price statistics. Arief Noor Rachman discusses the construction of a hedonic price index for residential properties in Jakarta based on asking prices scraped from web portals. The results are compared with existing indices based on appraisal data and are very promising. The methodology could serve as an example for countries that are developing similar indices but have no access to data on transactions or appraisals.

Karola Henn, Chris-Gabriel Islam, Patrick Schwind and Elisabeth Wieland use actual transaction data on package holidays to develop experimental consumer price indices for these services for Germany. They discuss the practical and methodological issues involved and how they can be resolved, and analyse a variety of different index formula. The paper thus makes an important contribution to the continuously expanding literature on the use of transaction data for price statistics.

I hope you enjoy reading these four important articles, which together demonstrate the continuous search for improved measurement methods in macroeconomic statistics, in order to serve user needs better.

Paul Konijn

Editor of EURONA

# 1 The measurement of public goods: lessons from 10 years of Atkinson in the United Kingdom

FRED FOXTON, JOE GRICE, RICHARD HEYS & JAMES LEWIS ([1])

**Abstract**: When considering issues of measuring welfare beyond gross domestic product (GDP), a key ongoing, but unfinished, agenda concerns how to measure the outputs of goods and services which are 'free at the point of delivery'. Public services such as schools and health services are major examples of this kind. Over a decade ago, Sir Tony Atkinson provided a principled framework for this end. Consistent with the basic principles of national accounting, he advocated an approach by which this output should be measured as the value added by the services concerned, where this equates to the incremental contribution to (monetised consumer utility from) outcomes resulting from the delivered outputs. This value, in turn, equated to the improvement in outcomes directly attributable to the activities of the public services concerned. Implementing this approach, as Atkinson recognised, is by no means straightforward, but the United Kingdom experience shows that considerable improvements can be made. Working with experts and practitioners, quantity and quality measures can be identified and used to give a good approximation of the value added by key public services, and thus their contribution to GDP. New data and intelligent use of existing data means this can be done at low cost and in a way which maximises stakeholder understanding and acceptance.

But national statistical institutes are also now grappling with a second task; measuring changes in welfare or well-being more generally, regardless of how they are generated. Health outcomes — for example, life expectancy or healthy life expectancy — are influenced by a variety of factors besides publicly-funded health services: diet, smoking prevalence and other lifestyle choices are obvious determinants. So, the central tasks under this agenda become firstly the identification of appropriate measures of outcome changes, secondly determining how much value our societies place on those changes, and thirdly to understand the relationship between the impact of the public service and other factors on the headline outcome measures.

**JEL codes**: C46, C82, E01, H4, H5, I1, I21, I38, K42, 047

**Keywords**: national accounts, GDP, public sector, services, economic well-being, quality adjustment, productivity

([1]) Office for National Statistics, United Kingdom. The views expressed within are the personal views of the authors and do not represent, or claim to represent, the views of the Office for National Statistics.

# 1. Introduction

Current debates about measuring the impact of the digital economy, (specifically free digital goods which deliver welfare gains to consumers), even if their exact treatment in the national accounts is under debate, need to be seen in the context of a larger group of transactions which are also free, or nearly free, to consumers; these are mainly public services. In the United Kingdom, around 20 % of gross domestic product (GDP) is accounted for by the output of public services. Other G7 countries exhibit similar magnitudes ranging from 19 % to 24 %, with the one exception being the United States at around 14 %. Measurement of these free goods is a common issue affecting almost all countries.

The United Kingdom has had an interest in this question since 2003 when the then National Statistician, Len Cook, asked Sir Tony Atkinson to conduct an independent review of the measurement of government output in the national accounts, with a final report produced in 2005. The resultant publication was a seminal text which informed the development of the System of National Accounts 2008 (2008 SNA) in how to conceptualise and then empirically measure the outputs of public services contained in GDP. The United Kingdom, alongside several other countries, pressed ahead with implementing these methods. This work managed to address the largest parts of the public services, but gaps remained.

The Bean Review (2016) commended the Office for National Statistics (ONS) for this work but identified that renewed efforts were needed to update the methods being applied where quality adjustments were in place and to create new adjustments where these were not.

This paper makes four contributions to this debate. First, it draws attention to the importance of these issues both in terms of economic activity and more widely to consumer welfare: whilst the impact of changes in digital technology over the last 20-30 years are important, considering the life-saving and life-enhancing improvements in medical care over the same period gives important context. Secondly, to re-iterate a commonly missed Atkinson recommendation: the fast pace of change in public service ([2]) delivery and usage means that methodologies need to be kept under regular monitoring and updated as required. Thirdly, the paper draws out key lessons the United Kingdom has learnt over this period which the authors hope might contribute to the process of mutual learning. Finally, it highlights how a better understanding of the public sector's contribution can only enhance efforts to measure economic welfare.

The paper is structured as follows:

- a brief account of the historical context of measuring public services in the United Kingdom, in the wake of the 1993 SNA, and the problems that were encountered that led to the Atkinson Review;
- an account of the Atkinson Review and its implementation in the United Kingdom;
- a discussion of the current methods used to calculate quality adjustments in the United Kingdom;
- a summary of the most significant issues identified in the United Kingdom in measuring public service outputs and outcomes, and how these have been addressed over the last decade;
- a discussion of the challenges in capturing welfare gains related to public services alongside other non-GDP welfare gains in any new metric; and,
- conclusions.

([2]) This paper focuses on 'public service' rather than the 'public sector' simply because mainly public services are now delivered via both the public and private sectors in many countries.

# 2. Measuring public service output and productivity: the historical context

The treatment and measurement of public service output and, by extension, public service productivity, has long been known to raise tricky but important issues. Quite clearly, their measurement is not straightforward. Most transactions included within GDP are measured at their market or exchange value. But most outputs provided by the public sector — health services or public provision of education, for example — are non-market services. So, while such services clearly have value, there is no observable price to guide the valuation. The value, therefore, must be imputed and this may not be simple to do ([3]).

The founding fathers of national accounting wrestled with how public service outputs should be treated in the accounts and indeed some, like Kuznets, proposed excluding them entirely. Hicks changed his mind at least twice on this question. In the event, the consensus was to adopt a convention — the so-called 'output equals inputs' convention — whereby these non-market outputs were deemed equal in value to the inputs used to produce them. The implication of 'outputs equals inputs' is that public service productivity is always constant, with its growth rate, by definition, zero.

Leaving aside the measurement complexities, there are important reasons for taking public service output and productivity seriously. One is the sheer scale of the transactions involved. In the United Kingdom, for example, non-market public service output accounts for around a fifth of GDP ([4]); the sector is over twice the size of manufacturing. So, omitting these public services from the national accounts would be to ignore a major part of the value which the economy generates. Similarly, to do so would be to overlook a material contribution to the overall productivity of the economy. Nor does such productivity performance simply mirror that of the rest of the economy. In recent years public service productivity in the United Kingdom has been rising while the productivity performance of the rest of the economy has been stagnant.

A second reason why public service productivity is important relates to fiscal policy. Finance ministries are continuously in the horns of a dilemma, though one whose acuteness varies over time. On the one hand, the political pressure for improved public services is strong. Citizens as users have rising expectations of what they receive from health services, from publicly provided education, by way of social care and so on — no less than they have rising expectations for economic performance overall. Where many public services are key to tackling inequality and improving life chances, as these issues are important in public debate and amenable to improved public services, understanding the output of the public sector helps users understand governments' steps to tackle inequality. But citizens as taxpayers are also reluctant to pay the rising taxes that might finance the improving public services. The only way to square this circle is to improve the efficiency and effectiveness of how taxpayers' funds are used, so that through increased productivity, more output is produced by the same amount of taxpayers' money.

([3]) This does not imply that where prices exist measurement is self-evidently simple. Capturing quality change and ensuring price deflators accurately compare like-for-like products are still substantial challenges even when prices exist. Whilst in this paper the authors predominantly reflect on the instance where prices cannot be observed, many of the issues described are still of relevance to countries where these services are delivered via the market.

([4]) This can vary marginally by year selected.

Accordingly, monitoring public service productivity is of policy importance over and above the sector's (sizeable) contribution to productivity performance overall.

Third, the performance and efficiency of public services conditions the productivity of the rest of the economy. A well performing legal system, for example, is vital for underpinning a well-functioning commercial sector. An efficient and well-performing health service is a major contributor to a healthy and productive workforce, while the outputs of publicly provided education make a direct contribution to the nation's human capital. Arguably, the same outputs also feed into social capital and thus again underpin a well-performing economy overall.

Given the importance of these issues for economic commentary and policymaking, the balance of opinion in the national accounting community increasingly moved towards thinking that the 'outputs equals inputs' convention was untenable. There was no reason to suppose that it gave an accurate view of how the outputs and productivity of this growing sector were behaving within the overall economy. Since, by definition, it implied necessarily unchanging productivity within the sector, it could give no useful information regarding the other two issues: how well public services were making use of taxpayers' funds or how productively public services condition the performance of the rest of the economy. These drawbacks from 'outputs equal inputs' were substantial.

Accordingly, the System of National Accounts 1993 (1993 SNA) recommended that, in future, countries should move away from the previous convention and instead adopt methodologies which measured the output of public services directly, using observable information relating to these services. This would mean of course that there was no reason why the estimated outputs from such methodologies would equate to the observed inputs. Consequently, it would also be possible to estimate how productivity in these various sectors was changing over time.

The ONS was one of the early movers, together with a handful of other national statistical institutes (NSIs), in taking forward this new agenda. By the late 1990s, measured by value, some two thirds of public service outputs were measured directly. The remaining third continued to be measured by 'outputs equals inputs'; the so-called collective services, particularly the defence sector, were the main part of this residuum ([5]). However, not long after the new methodologies were put in place, the estimated productivity series began to demonstrate paradoxical behaviour. Having been rising at fairly steady rates up to 1997, the estimated productivity of the directly measured sectors fell by over 20 % in the four or five years after 1997. It was hard to understand why the estimates were showing such declines. Nor was there any corroborating evidence to suggest that such declines had occurred. Accordingly, users' confidence in the validity of the estimates became increasingly strained. Since the output-driven estimates also now fed into the United Kingdom's overall national accounts, confidence in those, too, was also in question.

In these circumstances, at the end of 2003, the then United Kingdom National Statistician, Len Cook, asked Sir Tony Atkinson to conduct an independent review of methodologies to measure public service output and productivity. His terms of reference also included looking at the way the ONS had approached the new SNA agenda and its implementation of direct measurement methodologies. The Atkinson Review lasted for just over a year and Sir Tony published a report in January 2005 setting out his conclusions.

([5]) This does rule out that the same cash amount of public expenditure can result in higher outputs. If public authorities can buy the relevant inputs more cheaply, then the same amount of cash will buy a higher volume of inputs and thus under the 'output equals inputs' convention be deemed to generate higher output. But this is an effect from more efficient procurement and should be distinguished from the productivity channel.

## The Atkinson Review and its legacy

The Atkinson Review was a milestone in this agenda. The report clarified many issues and through its recommendations proposed a model for measuring public service outputs including a research and implementation programme in the main public service areas. Len Cook accepted Atkinson's conclusions, subject to underlining that their full implementation would take time and be conditioned by availability of resources ([6]).

Fundamentally, Atkinson agreed wholeheartedly that the SNA had been right to counsel direct measurement of non-market public services. The drawbacks of the traditional 'outputs equal inputs' convention were too great to be acceptable, for the reasons set out earlier in this paper. By the same token, the ONS had been right to take up this agenda. The issues observed in the United Kingdom data were real ones but were rooted in how the agenda had been implemented, as discussed further below, not because the overall agenda was problematic.

Atkinson's report saw the problem as being the ONS's failure to base its methodologies and estimates on a clear set of explicit principles. Not unnaturally, when faced with a difficult task, in many cases ONS statisticians had sometimes used stop-gap methodologies and/or readily available indicators or other data sources, in the hope that this would be better than nothing, but these did not necessarily relate directly to what was needed to measure public service outputs. Experience showed these hopes were not always realised: it can be argued, in some cases, that the procedures had led to estimates which were worse than not having anything.

The complete set of Atkinson's principles is shown in Annex A. One superficial reaction to them is that many look like common sense. Who would not be able to agree to them? On the other hand, their usefulness and power comes from employing them as a yardstick against which to compare the actual procedures which were in place. They quickly highlighted areas where the ONS's existing procedures did not measure up. This gave a clear indication of where remedial action was required as well as helping guide the nature of the remedial action and revised procedures.

One particularly important principle related to what should, in theory, be included in a country's national accounts and therefore what the methodologies should be striving to capture. Atkinson contended that the key consideration in national accounts was value; thus, GDP could be considered as the cumulative value added from the economy, going through the various stages of production. It was therefore essential to avoid measuring public service output solely by what were essentially activities — say, the number of medical procedures performed or the number of pupils taught, particularly where such measures may incentivise perverse outcomes; such as fire-protection services being measured using the number of fires they put out, where increasing fire protection activity would lead to a reduction in output, rather than a growth.

([6]) The Atkinson approach, measuring the incremental contribution to consumer utility as the measurement principle, is not the only approach which could be taken. The main alternative, a measurement target that adopts a producer perspective is described in, for instance, Diewert (2017) and Schreyer (2012).

The problem he saw was that such activities may or may not have value. His private sector analogy was the production of broken bricks. A factory which produced only broken bricks would find its output next to negligible since the broken bricks would have little or no value, as opposed to well-produced whole bricks which, of course, would have value. In the public services, the equivalent issue was to establish whether the hospital procedures carried out or the number of pupils taught were adding value or otherwise; what was the quality of the 'bricks' they represented.

A key principle was therefore that the estimates of public service output should be quality adjusted, to reflect the incremental contribution to (monetised consumer utility from) outcomes resulting from the delivered outputs. At a common-sense level, the value of a health care intervention clearly depends upon its quality. The procedure is of value only to the extent that it leads to a health outcome superior to a counterfactual where the procedure had not been carried out. This leads inevitably to the question as to how outcomes should relate to the estimates. Traditionally, national accountants had been reluctant to consider outcomes as relevant and with some good reason. In most countries, life expectancies and healthy life expectancies have risen significantly over time. While improving health services have played a part in this, the broad evidence is that this has been a minority contributor with factors such as improving diets, falling levels of smoking and healthier environments being much more important. It would therefore be quite wrong to ascribe the whole value of the improved health outcomes to the output of healthcare sectors. On the other hand, to the extent that an improved health outcome can be directly attributed to the activities of healthcare systems, then that should be taken into account in the estimated output.

The Atkinson Report was widely debated in the years following its publication. Its approach was largely accepted and helped shape the revised System of National Accounts 2008 (2008 SNA). The principle of allowing for quality adjustments in estimates of output was accepted and emphasised, as part of a wider trend of economists becoming increasingly comfortable in addressing social welfare function issues. The European System of Accounts (ESA) which generally follows the SNA, as its guiding principles, surprisingly, and somewhat regrettably, took a flatly opposite view and banned quality adjustments within its 2010 iteration, focusing on, for example the quantity output method for individual non-market services such as education and health ([7]).

This illustrates how contentious this topic remains. The decision was purportedly in the interests of international comparability but the authors would argue there seems to have been some muddled thinking at work. Imposing arbitrary comparability in methods does not necessarily serve the interests of comparability of the realities. With quality adjustments not allowed, those countries where public services have improved in quality are estimated with outputs below the reality and conversely for those where quality improvement has been relatively low. This can only prejudice rather than help international comparability. Eurostat has been organising work to review this issue so, hopefully, this is on the way to being resolved.

([7]) ESA 2010 (§10.29-10.30).

Since the Atkinson Review, the United Kingdom has delivered public service output and productivity estimates, with varying degrees of success, differing both between and within service areas as shown in Figure 1. The approaches are categorised broadly into three types.

**'Output equals inputs'** — accounting for around 38 % of public service output in 2016, this approach assumes that the volume of output is equivalent to the volume of inputs used to create them. Typically capturing what are referred to as 'collective services' (such as defence), this convention is used when the output of a service area is conceptually difficult to define and/or measure. As a result, productivity is assumed to remain constant and growth will always be zero. This is the least satisfactory method.

**Quantity output** — representing 12 % of public service output in 2016, this approach, in line with that recommended in ESA 2010, uses long-standing indicators of activities known as cost-weighted activity indices (CWAI). Here an index is constructed as the weighted sum of change in the level of different activities from one year to the next. As most public services do not have a market price to use as a weight, given they are not sold on a market, the costs of producing a unit of activity (unit cost) are used as a proxy. Although this cost weighting occurs, the use of measured outputs is believed to be an improvement on the previous input-based methodology and is used as a measure of output in United Kingdom estimates of total public service output. More detail about this approach, and the steps involved, can be found in Annex B.

It is, however, recognised as the second-best approach. While some elements of quality change can be captured (for example, through the differentiation of activities), a CWAI will fail to capture all quality improvements. An example of this would be the gradual introduction of a new healthcare procedure which yields better outcomes at lower cost. If activity growth is connected between old and new procedures, the CWAI approach would result in the weight of this activity in the index gradually falling over time. If the old and new procedure are recorded separately in the index, the CWAI approach would even lead to a fall in output.

**Quality adjusted output** — the third category then accounts for the remaining 50 % of public service output in 2016. This approach takes the CWAI as a starting point and builds on this by adjusting the quantity output to take account of changes in quality, in line with the recommendations of the Atkinson Review, reflecting improvements in outcomes that can be attributed directly to public service activity ([8]).

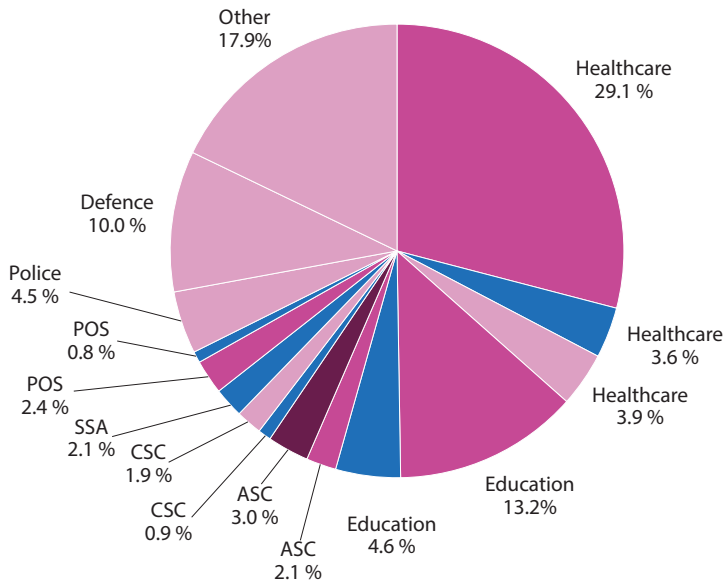
Within the market sector, higher-quality variants of outputs can be picked out. As higher-quality outputs sell for more than lower-quality outputs, the change in quality is accounted for by the price differential. This is, of course, much harder (but no less important) to do for public services because service users do not pay directly for services, and thus there is no user-driven differential to use. Where they can be identified quality metrics are, therefore, used to augment volume data, based on how far outcomes can be attributed to public services, to give a well-based measure of the public service output concerned.

([8]) In relation to adult social care we have quality adjusted output, where output is calculated on an inputs = outputs basis. We have included this in 'quality adjusted output' for the sums in the text.

It is important to note that such quality adjustments are explicitly excluded from the measurement of output in the national accounts central framework by ESA 2010, and are not part of the output series used in other ONS measures of productivity.

More details about the general methodology can be found in Annex C, while more specific details are provided later in Part 3 of this paper.

**Figure 1: Output-type share by service area, United Kingdom, 2016**




Note: shares may not sum to 100 % due to rounding. POS (public order and safety) includes courts and probation services, the prison service and fire-protection services. Other government services includes services such as economic affairs, recreation and housing. ASC refers to adult social care. CSC refers to children's social care. SSA refers to the social security administration.

*Source:* Office for National Statistics — Public service productivity: total, United Kingdom, 2016 (ONS (2019b))

Atkinson made a further important recommendation to 'triangulate' estimates with corroborating evidence when assessing public service output. Such evidence might be subjective or objective. The sharp downturn in the ONS' estimates of public service output that had led to the Atkinson Review being set-up turned out to be largely illusory and due to problems with data sources and methodologies. A subjective source of evidence that might have shown up the problems earlier would have been talking to practitioners and expert commentators. When, during the Review, they were asked what might have caused the sharp downturn in productivity, their invariable response was that they were not aware there had been such a downturn.

Such evidence would not have been conclusive but would, at least, have rung alarm bells. Such subjective evidence can, moreover, be supplemented by objective evidence. In the hospital sector, one of the key factors affecting efficiency is the average length of stay. With relatively fixed hospital capacity in the short-term, a shorter length of stay allows more patients to be treated. So, if there had been a sharp downturn in health sector productivity, it would have been reasonable to expect the average length of stay to have increased. But, in fact, it had not behaved out of the ordinary: again, not conclusive but another possible alarm bell to give reason to question the estimates.

A third precept from the Atkinson Review stemmed from the nature of what it saw to be the task. Assessing changes in the quality of various public services, or for that matter collecting and assessing triangulation evidence, may well not be within most NSI's core competences.

Fortunately, such issues are the core business of other communities. Assessing, for example, the quality of teaching and the contribution that schools make is what many education experts spend much of their time doing. Similarly, such issues in the healthcare sector are a central preoccupation of public health experts, epidemiologists, health economists and so on. Practitioners, by definition, are a complementary source of expertise. So, too, are government departments and other public authorities, who will have much greater expertise and experience of the services in their fields than an NSI could ever hope to muster. Atkinson therefore recommended that the ONS (and other NSIs) should form networks with such experts to allow them to tap into the expertise that would be needed to compile authoritative estimates of output in public services.

One important purpose that such networks could serve would be to feed into periodic reviews of the ways that public services are delivered and whether intervening changes mean that the original data sources and methodologies for compiling output and productivity estimates remain valid or whether changes are necessary. Models for delivering public services change no less quickly than the business models underpinning private sector activity. So, without such periodic reviews, there would be the possibility of maintaining methodologies that no longer corresponded to the real world. Of course, in principle, detecting such changes should be an ongoing concern. But, periodic reviews should serve as a safety net to ensure that relevant changes are picked up.

The Atkinson Report set out a principled approach which it recommended as a general model for measuring public service output and productivity. It also had chapters with suggested agendas for applying the approach in four key areas:

- healthcare services;
- public sector education services;
- public order and safety (specifically the criminal justice system);
- services relating to adult social care.

It recognised that completion of the work programme to fulfil these agendas would take several years. The rest of this paper discusses the United Kingdom's experience in taking this work forward and some of the principal lessons learned. Then, in light of this work, it considers both the welfare implications of public services but also how we treat those welfare gains which are not attributable to changes in public service provision.

# 3. Current methods of calculating Atkinson quality adjustments

We begin by briefly explaining the methods used to derive the four quality adjustments currently in use in the United Kingdom. It is important at this point to note that these all apply to services which are received by an individual (one person receives an operation, one individual receives any particular qualification, and so on), as opposed to collective services, such as defence. Collective goods, which are non-excludable by nature, present a further set of challenges in terms of measurement over and above those described below, which focus on the individual.

## Healthcare

In the United Kingdom, health care is primarily a public service under the Atkinson Review definitions. Nearly 80 % of the United Kingdom's health care expenditure is publicly-funded with much of this public expenditure funding free-at-the-point-of-use care through the National Health Service (NHS) ([9]).

The task of valuing the output and measuring the productivity of a free-at-the-point-of-use service, without insurers or other intermediaries negotiating prices from care providers, therefore faces the same challenges Atkinson sought to address across other public services.

But mitigating the considerable challenge of measuring the productivity of a service for which a price does not exist, the NHS provides the advantage of a wealth of data, collected on a uniform basis from all NHS providers.

### MEASURING HEALTH CARE OUTPUT

As with other public service sectors, quantity output is measured through a cost-weighted activity index (CWAI) ([10]). The data for this comes from detailed published management information. NHS provider organisations responsible for hospital, community and mental health care report detailed data on activity and unit costs as part of the process of setting reimbursement rates for the thousands of different activity types carried out across these sectors, as well as for use as a management information resource. For this purpose, activity and expenditure are analysed by healthcare resource group (HRG) and by care setting. The HRG system provides a more detailed and precise treatment-classification system as an alternative to the internationally-used diagnosis-related group (DRG) system, with over 25 000 individual activity types in the most recent years.

([9]) The NHS provides healthcare in Great Britain. In Northern Ireland, the Health and Social Care Service provides similar free-at-the-point-of-use care. For brevity, 'NHS' is used to describe all public health care services in the United Kingdom.

([10]) Produced by chain-linked Laspeyres indices. Estimates of quality adjusted output are produced in a similar manner as explained in Annex B.

For other elements of publicly-funded health care outside of NHS hospital, community and mental health care provision and drug prescriptions, data are scarcer. Data availability is particularly problematic for general practice, where output is currently measured using modelled estimates based on historical and demographic data, and for the rapidly growing component of NHS-funded services that are outsourced to independent sector providers.

NHS hospital, community, ambulance and mental health provision accounts for 64 % of total spending according to the most recent data, with a further 10 % from prescription drugs. Other family health services, of which general practice is the largest component, along with the more easily measurable dental and ophthalmological services, account for 15 % and services purchased from non-NHS providers a further 11 %.

The United Kingdom's public service health care output therefore combines a large element of some of the most precise output measures available for United Kingdom public services, with estimations needed for some of the other elements of the service. But as with other service sectors, the limits of cost-weighted activity in determining the value of public services provided still hold across all service elements. Hence a quality adjustment is required.

## MEASURING HEALTH CARE PERFORMANCE AND OUTCOMES

The comparative wealth of data available for health care extends to data on the quality of services. Here, a large variety of measures are available — NHS performance statistics provide monthly measures of performance against targets for a range of activities, while outcomes data from life expectancy to cancer survival rates provide indicators of the ultimate goals of the health service.

However, this trove of data does not automatically translate into the quality adjustments envisaged by Atkinson for output and productivity.

Consider the use of NHS performance indicators as quality adjustments and, as an example, accident and emergency (A&E) department waiting times, which are one of the NHS's highest-profile headline performance measures.

We can track the percentage of A&E patients who are seen within the NHS's national four-hour waiting time target. But it is not clear how a change in a quality adjustment incorporating the proportion of patients seen in four hours should affect the value of A&E output. Should we give equivalence to the volume and quality measure such that a 1 % increase in activity and a 1 % decrease in patients seen within the time target are roughly the same as a stable value of output?

This would imply that the value of providing A&E services to patients after the four-hour target is near-zero. However, given that patients counted in the activity data after a wait of four hours have endured the loss of their valuable time in surroundings not of their choice to receive care, it appears clear, even to a logic-seeking economist, that the value patients place on receiving emergency medicine services is greater than zero. So, such a simple solution would clearly be inadequate.

And the problems of how to apply such a performance measure to output do not stop with only the question of how such expenditure is scaled. Such a performance indicator only reflects one aspect of quality and research shows a tendency of providers to modify their behaviour to meet the minimum requirements, but not necessarily the spirit, of performance targets (Kings Fund (2017)). For instance, the four-hour waiting time target may encourage A&E departments to prioritise seeing patients who are approaching the four-hour mark, but improvements in performance against the four-hour target may not reflect shorter waiting times for patients in other parts of the waiting times distribution.

Therefore, robust quality adjustments cannot simply be drawn from the NHS performance targets. Instead, they should inform the effect of health provision on the outcomes they are trying to achieve.

One alternative measure from the health economics literature provides a conceptual framework which fits the criteria for quality adjustment far more closely, the quality-adjusted life year (QALY). The QALY is a tool for evaluating health care interventions that was first developed in the 1960s and 1970s and is now used globally (Mackillop and Sheard (2018)). The QALY is particularly prominent in the evaluation of health care in the United Kingdom, where the National Institute of Clinical Excellence (NICE) uses it to make recommendations on what treatments should be funded on the NHS.

While there is no single definition of a QALY, NICE uses the definition that a QALY is *a measure of the state of health of a person or group in which the benefits, in terms of length of life, are adjusted to reflect the quality of life.*

The QALY thus has two elements, a health-related quality of life element and a temporal element; and can therefore combine the effect of improvements in health-related quality of life and increases in the length of life resulting from treatment.

Health-related quality of life is measured on a scale between zero and one, with zero being a state equivalent to death and one representing perfect health. For the evaluation of health care, the gain in health-related quality of life from an intervention is then measured across time to produce a measure of QALY gain, such that a gain of one QALY represents one additional year of life in perfect health following the intervention.

However, while the QALY serves to provide much of the theoretical grounding for a quality adjustment, the quality adjustment used by the ONS cannot simply consist of a change in measured QALY both for the practical reasons that no systematic regular data collection on patients' health-related quality of life before and after treatment exists, nor are consistent data available on the increase in patients' life expectancy resulting from treatment, but also for the conceptual reason that changes in health states are not just caused by health provision, but also by an array of other factors.

## QUALITY ADJUSTING HEALTH CARE OUTPUT

Unlike other adjustments where the ONS and the relevant government department generally undertook the relevant work, given the challenges of constructing a quality adjustment to meet the principles from the Atkinson Review, the current health care quality adjustment was designed through a rigorous process, which set out the measurement framework, involving an expert group of health economists ([11]).

The construction of the measure incorporated a range of relevant factors, while taking care to minimise combining metrics which would overlap and thus record the same quality drivers multiple times. For instance, the elective inpatient care adjustment combines health gain, survival, waiting times and patient satisfaction, thereby covering the main aspects of care quality. The quality adjustment used by the ONS for healthcare output continues to be based on this research.

The measure produced can be divided into three components:

- hospital procedures adjustment;
- primary care outcomes adjustment;
- patient experience adjustment.

While we will discuss each of these in turn, of the three, the hospital procedures adjustment is by a large margin the most significant in terms of its effect on the measure, while also being by far the most complex. The hospital procedures adjustment continues to be produced by the Centre for Health Economics at the University of York and is used both to quality adjust healthcare output in the ONS measure and for a separate productivity analysis carried out by the Centre for Health Economics (see Dawson et al. (2005)).

### *Hospital procedures adjustment*

The quality adjustment utilises the hospital episode statistics (HES) dataset, whilst incorporating other data from various sources. The HES dataset is a highly detailed administrative dataset recording details of all patients receiving hospital treatment from the NHS in England. Observations in HES are coded to appropriate activity types using the aforementioned HRG system which is used to produce cost-weighted output.

The team did not try to value welfare gains on a QALY basis, although the concept of QALY is central to the hospital procedures quality adjustment, for the following reasons:

- There is no certain value to one QALY — NICE does not specify a single value, though their treatment recommendations imply a value for one QALY of GBP 20 000-30 000. However, other bodies use different approaches to value health, with the Department for Transport using a single figure per life lost to evaluate road safety interventions (Glover and Henderson, 2010).

([11]) Funded by the United Kingdom's Department of Health, the project team consisted of several economists from the Centre for Health Economics at the University of York and the National Institute of Economic and Social Research (NIESR), along with the input of other involved bodies, including the Department of Health and ONS.

- The value of a QALY should vary over time with average incomes and the marginal utility of income, but an increase in incomes should not be attributed to a quality adjustment for the NHS. However, holding the value of a QALY constant could result in the quality adjustment effect declining over time as cost inflation affects the ratio between quality and quantity value.
- Other factors beyond pure QALY gains may be important for the quality adjustment, such as patient experience.
- As discussed in Part 4, below, QALY gains could be influenced by other factors outside of the influence of the NHS, such as the output of other services, environmental factors or changes in patients' behaviour.

The hospital procedures adjustment which was instead developed can itself be broken down into three sub-components:

- estimate of health gain;
- short-term survival;
- waiting times.

The health gain estimate is an attempt to derive a proxy for the gain in health-related quality of life on an equivalent zero-to-one scale as is used in the calculations for QALYs. While, as previously mentioned, there is no systematic collection of health-related quality of life for all patients across the NHS, patient reported outcome measures (PROMs) are collected from patients across two high-volume treatment groups — hip replacements and knee replacements, and until 2017, were also collected for groin hernia and varicose veins procedures. The PROMs give measures for health-related quality of life gains before and after treatment using the EQ-5D scale, a widely used assessment framework — which uses patient responses to questions on ability to pursue usual activities, anxiety/depression, pain, mobility and ability to self-care — to produce a health-related quality of life score on a scale of zero to one.

However, these PROMS measures cover a tiny fraction of the total number of patients receiving hospital treatment on the NHS. The health-related quality of life gain for the majority of patients therefore needs to be estimated. The Centre for Health Economics produced a single 'rough estimate' ([12]) for all remaining elective treatments (procedures scheduled in advance) and a single 'rough estimate' for all non-elective treatments (urgent, unscheduled procedures). These estimates assume a greater health gain for non-elective procedures as patients generally arrive in a worse health state than elective patients, and so experience a greater health gain ([13]).

The gain in health-related quality of life is spread across remaining life expectancy as derived from the ONS data on life expectancy by age, although due to a lack of data, no adjustment is made to counter the effect that treatment may extend life expectancy or that patients may have a lower life expectancy than the general population due to their pre-existing health issues. Gains in health-related quality of life across the future are discounted using the social time preference rate.

([12]) The Centre for Health Economics' assessment, not that of the authors.
([13]) Some further adjustments to these rates are applied to procedures with high mortality rates to avoid the quality adjustment giving a negative valuation to these procedures, but for brevity we will not explore the details of these further adjustments here.

Post-operative short-term survival rates are then incorporated in this measure to reflect health-related quality of life falling to zero for patients who do not survive the procedure or die before being discharged.

A waiting times factor then incorporates the forgone potential health gain for treatment being delayed, with waiting times at the 80th percentile taken to reflect the importance of uncertainty and the risk of long waiting times to patient well-being.

### *Application of the hospital procedures adjustment*

The overall adjustment for hospital procedures is then calculated from the change in the health-related quality of life element (health gain/survival rate factor) multiplied by the change in the temporal element (life expectancy/waiting time factor). This adjustment is calculated individually for each HRG and then applied to the output data which is also calculated at the HRG level, meaning the quality adjustment is not simply applied as an aggregate of all procedures carried out, but incorporates the same cost-weighting as non-quality adjusted output.

However, the complexity of the calculation means that the drivers of this quality adjustment can be difficult to discern and the direction of effect not always immediately intuitive. Table 1 explains the effect of these changes.

**Table 1: Effect of changes in components of hospital procedures quality adjustment on output**

| An increase in … | Effect on quality adjusted output | Mechanism of effect |
|---|---|---|
| Health-related quality of life (HRQoL) gains reported in patient reported outcome measures | Increases | Changes HRQoL gain |
| Proportion of treatments that are elective ([1]) | Decreases | Changes HRQoL gain |
| Post-operative survival rates | Increases | Changes HRQoL gain |
| Average age of patients being treated *(life expectancy at birth unchanged)* | Decreases | Changes length of period over which the gain is experienced |
| Life expectancy at birth *(age at treatment unchanged)* | Increases | Changes length of period over which the gain is experienced |
| Waiting times (80th percentile) | Decreases | Changes length of period over which the gain is experienced |

([1]) The Centre for Health Economics assigns a greater health gain factor to non-elective treatments than elective treatments (see paragraph on pp. 13).

***Primary care output adjustment***

While the hospital procedures adjustment provides a quality adjustment for a large proportion of spending, the NHS also comprises many other smaller services. As previously mentioned, the relative paucity of detailed output data for other NHS services also applies to quality data.

For general practice, the largest component of primary care, a quality adjustment has been built using a selection of appropriate outcomes data from the quality and outcomes framework (QOF), an incentive scheme for general practitioners (GPs). These measures relate to the extent to which patients' health risk factors fall above or below risk thresholds, thus incentivising GPs to monitor, medicate and promote behaviours for healthy outcomes; for example, the proportion of patients with coronary heart disease who have blood pressure and cholesterol readings above a threshold. The quality adjustment is scaled down to reflect the fact that only a small proportion of the population has the relevant risk factors.

The GP outcomes adjustment is a demonstration of the fact that the collection of quality and outcomes data can vary as policy changes. Data from the QOF improved rapidly after their introduction as an incentive scheme in the early 2000s, but the scale of improvements decreased in subsequent years as GPs moved closer to the maximum achievable scores. As the QOF system has matured and the gains in outcomes have become more marginal, the number of measures collected has fallen, further reducing the proportion of GP activity that the quality adjustment covers.

***Patient experience adjustment***

The patient experience adjustment covers a range of NHS services and was included to account for the issue that the other aspects of the quality adjustment do not incorporate non-clinical aspects of care quality which may also be valued by patients, such as being well-informed and involved in decision-making, having good relationships with staff and having a comfortable environment.

The patient experience adjustment is calculated using data from the overall patient experience scores, which are based on national surveys covering in-patients, out-patients, emergency care, community mental health services and primary care, although the patient survey for primary care has been discontinued since the quality adjustment was designed.

Generally, the patient experience scores demonstrate only minor variations over time, and as with the GP outcomes adjustment, result in a relatively minor effect on the overall quality adjustment.

## Education

Education services comprise eight publicly-funded sectors ranging from pre-school to higher education training of teachers ([14]) and is captured using data on student numbers for the respective sectors. Unsurprisingly then, the two largest sectors are primary schools and secondary schools (including academies) both in terms of spend and active student numbers. Like other individual services, the output of education services is measured directly, reflecting changes in the aggregated activities delivered. However, looking deeper into this, the demography of the United Kingdom will mean that using pupil numbers alone gives only a very low or constant growth in education output over time, implying the volume of output of the public education system may not have significantly improved during this period.

This obviously abstracts away from quality factors such as the quality of teaching pupils receive; the depth to which a syllabus is taught; the individual attention afforded to them by teachers; and the skill sets developed. Therefore, if these changes in quality are accounted for, it would be expected that the volume of education service output would change, even if demographic factors hold student numbers constant. Therefore, following Atkinson (2005), additional steps were incorporated to explicitly account for changes in the quality of provision in estimates of education output.

In general, the output of education sectors is quality adjusted in two stages. Firstly, student numbers are adjusted by attendance rates. In line with specific recommendations outlined by Atkinson, rather than using pure registered pupil numbers, adjusting by absence aims to provide a more accurate measure of the amount of teaching activity received by pupils, so absence (both authorised and unauthorised) are captured.

Secondly, metrics of 'high-level' attainment, using information about examination results, measure changes in the overall quality of services provided.

In Foxton (2018a), output associated with both primary and secondary schools is adjusted using the average point score (APS) per student at the General Certificate of Secondary Education (GCSE) level or equivalent examinations, which are normally taken during the student's 11th year of schooling. It is the best current measure for the annual change in the quality of output. It rests on the assumptions that the change in the APS can be used to approximate quality, and:

- should be applied to all pupils in primary and secondary schools ([15]) (from reception class to the end of the sixth form) in the United Kingdom;
- is an adequate approximation for all educational outcomes, for example attainment after the age of 16 and development of wider outcomes such as citizenship.

([14]) Initial teacher training (ITT).
([15]) Including Academies and City Technology Colleges (CTCs) in England.

As these examinations vary across geographical areas, the APS quality adjustment is applied to primary and secondary school output in each country separately. The APS at GCSE level for England and Wales are provided by the Department for Education and the Welsh Government respectively, while the APS associated with Standard exams in Scotland are provided by the Scottish Government. For reasons of data comparability and availability, the level of education quantity in primary and secondary schools in Northern Ireland is quality adjusted using the APS of English schools. Initial teacher training (ITT) quantity in each country of the United Kingdom is adjusted using the QTS award rate for England, which is also provided by the Department for Education. Here the implicit assumption is made that changes in quality in ITT in Wales, Scotland and Northern Ireland follow the trend in England ([16]). This and a number of other factors in relation to the measurement of education continue to undergo revision in the United Kingdom.

## The criminal justice system

Introduced as part of Foxton (2018b), when measuring the output and productivity of public order and safety (POS), explicit adjustments are made to the measure of output from the criminal justice system (CJS) to take account of changes in quality and improvements in associated outcomes. The basic activity measures, common to both public service productivity estimates and national accounts, consist of cost-weighted aggregates of services provided (such as prison bed-days or cases processed per court) which are paid for by the United Kingdom government. This is covered in greater detail both in Part 2 of this paper and Annex B. The quality adjustments applied then consider some of the aspects of quality not already captured by the simple activity measure of output for POS.

Within the POS service area there are four main components: fire-protection services, courts (which itself has five further sub-components), probation services and prisons. The quality adjustments are applied to a subset of these components, as shown in Table 2, which are identified as forming part of the CJS, alongside an indication of the weightings used. A quality adjustment is not applied to fire-protection services or County Courts, which deal with civil cases ([17]).

The criminal justice quality adjustment has four components:

- recidivism (re-offending) adjustment;
- prison safety adjustment;
- custody escapes adjustment;
- courts' timeliness adjustment.

([16]) This is a key issue in relation to geographical comparability — is it better to quality adjust all geographies, even when no quality adjustment data are available in that area, or is it better to present unadjusted data in these areas, even if this introduces a variation of its own.

([17]) United Kingdom court cases are divided into 'civil' and 'criminal'. Civil cases covering areas of family and contract law do not address 'criminal offences' and are therefore out of scope of a quality adjustment designed around addressing criminal behaviour. Similarly, fire-protection services are exempted from the described quality adjustment.

The first relates to achieving an overall outcome — reducing re-offending — for the whole CJS and therefore treats the CJS as one interlinking system that allocates and provides appropriate disposals ([18]) and rehabilitation services. It can, however, be argued that the associated sub-components may have specific target outcomes, in addition to reducing recidivism. Therefore, the remaining three adjustments relate to specific target outcomes for sub-components of the CJS.

**Table 2: Quality adjustment weights by output component**
(%)

| Component | Quality adjusted | Recidivism | Prison safety | Custody escapes | Courts' timeliness |
|---|---|---|---|---|---|
| Fire-protection services | No | | | | |
| Magistrates Courts ([1]) | Yes | 50.0 | | | 50.0 |
| County Courts ([1]) | No | | | | |
| Crown Courts ([1]) | Yes | 50.0 | | | 50.0 |
| Crown Prosecution Service ([1]) | Yes | 100.0 | | | |
| Legal Aid ([1]) | Yes | 100.0 | | | |
| Probation services | Yes | 100.0 | | | |
| Prisons ([2]) | Yes | 29.2 | 37.5 | 33.3 | |

([1]) Subcomponent of United Kingdom courts and related activities.
([2]) Weights for prisons quality adjustments are taken from prison and probation performance statistics 2014-2015. Further details on sources and methods used can be found in Foxton (2018b).

*Source:* Foxton (2018b)

### *Recidivism adjustment*

The recidivism adjustment is applied across all output associated with the CJS. It approximates the effect the CJS has on reducing the volume and severity of further crimes being committed by those who have gone through it — this being an important social outcome for the system. The ONS measure works by adjusting the cost-weighted activity indices of the service areas identified in Table 2 by a severity-adjusted rate of recidivism.

This adjustment itself is composed of three parts, the first being the change in the number of proven re-offences committed by adult and juvenile offenders categorised between crime types. These include such categories as violence against the person, robbery and fraud. Secondly, an adjustment is made to offenders, to account for differences between cohort characteristics and their likelihood to re-offend. The final adjustment made provides a weighting by which to aggregate together all re-offences. This weighting is based upon the relative severity of the re-offence and is derived from ONS (2016). More information on this source, as well as others used, can be found in Foxton (2018b).

([18]) A disposal can be thought of as an appropriate sentence for the crime and mitigating factors, such as repetition, aggravation, or factors which make the case more severe (for instance, assault with a weapon, as opposed to assault without a weapon).

### *Prisons safety adjustment*

The prisons safety adjustment relates to the number of incidents of assaults, self-harm and deaths that occur in prison custody. The purpose of this being to reflect that safety of prisons is an important component of the quality in the activity and services provided, as set out in the Prison Safety and Reform White Paper (Ministry of Justice (2016)).

We measure the number of incidents per 1 000 prisoners, which are grouped into 'severe', 'less severe' and 'those resulting in a death'. These groups are subsequently weighted and aggregated together based on their relative cost. This is achieved by using the total cost to society of workplace injuries as a proxy, taken from the Health and Safety Executive ([19]).

### *Custody escapes adjustment*

The escape adjustment relates to ensuring prisons fulfil the role of public protection and is applied to activities used to measure the output of the prison service.

The measure is based on changes in the difference between the number of escapes and a baseline of 0.05 % of the England and Wales prison population — a historic target used by the Ministry of Justice. The purpose of this being that as the absolute number of escapes approaches zero, the relative change year-on-year would have a disproportionate effect on a non-baselined quality adjustment index.

### *Courts' timeliness adjustment*

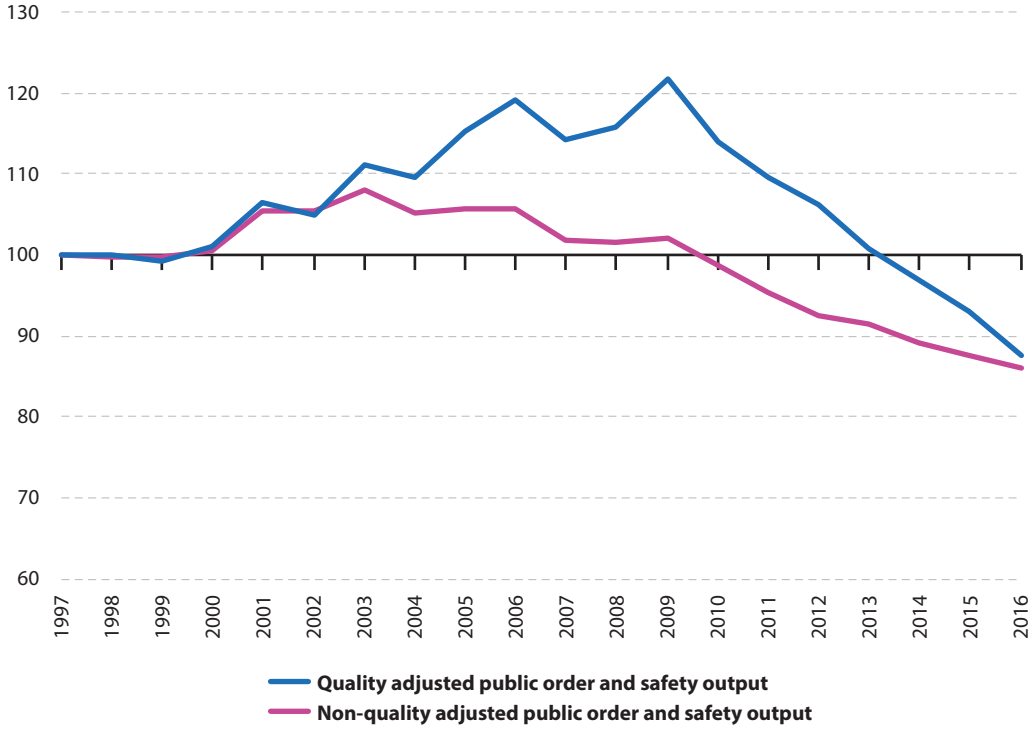
The courts' timeliness adjustment relates to the average time taken for criminal cases to be taken to completion, on the basis that the delivery of a sentence in a timely manner is favourable. However, there is currently no adjustment made to reflect whether there has been fair treatment of the suspect or victims or to allow the appropriate time for preparations of criminal cases with differing levels of severity or complexity.

For Magistrate Courts, the measure is based on the mean average time of charge and laying of information to completion. For Crown Courts, the measure captures the average waiting times experienced by all defendants and the mean time from main hearing to completion. As implemented, the measure accounts for changes in the average time taken to completion by criminal courts because increases in volume may reflect a worsening.

The net effect of this measure can be seen in Figure 2 which demonstrates the impact of quality adjustment on the output of the public order and safety sector in recent years.

([19]) This method is currently under review: the weight applied to a lost life compared with other injuries mean these data fundamentally drive this component. We are exploring with experts whether greater weight should be applied to other injuries which form the bulk of instances.

**Figure 2:** Non-quality adjusted output and quality adjusted output for public order and safety, United Kingdom, 1997-2016




*Source:* Office for National Statistics — Public service productivity: total, United Kingdom, 2016 (ONS (2019b))

## Adult social care

Adult social care (ASC) services comprise care and support provided to older people, adults with learning or physical disabilities, adults with mental health problems, drug and alcohol misusers, and carers. By spending, the largest two client groups are older adults and adults with learning disabilities. The services covered by ASC include placements in residential and nursing care homes, home visits by carers, day care services and supported living arrangements in accommodation adapted to users' needs.

Unlike the NHS, which has provided free-at-the-point-of-use health care to patients since it was formed 70 years ago, ASC services have not undergone the same funding and policy unification. While the NHS is a single public body in each of England, Scotland and Wales, responsibility for the provision of ASC services remains with local authorities ([20]). The funding of ASC comprises a number of streams, with the main sources being:

- Local authorities' own funds.
- Fees charged to clients, which for many services is subject to means testing based on clients' wealth and income.
- Payments from the NHS. There are a number of schemes under which NHS bodies transfer funds to ASC services. The funding transferred through these schemes has grown in recent years at the behest of government policy. The transfers are intended to financially support local authorities to relieve pressures on the NHS by reducing unnecessary hospital attendances by care clients (so-called 'bed-blocking') and promote co-operation between NHS and ASC service providers.

The provision arrangements are further complicated by most ASC services being contracted out by local authorities to the independent sector (typically private firms and charities), while a minority of services continue to be provided directly by local authorities.

There is also a substantial private sector, where clients purchase ASC services directly from private providers without necessarily involving local authorities.

As explained above, following the Atkinson Review guidelines, the remit of public service output and productivity is delineated by public spending as opposed to public provision. Therefore, the proportion of ASC services funded by local authorities and payments from the NHS is within public service output and productivity, whether provided by local authority or independent sector providers; while client-funded activity is excluded.

***Measuring the output of ASC services***

Activity and expenditure for ASC services are measured using data collected by the NHS from local authorities, enabling the construction of a cost-weighted activity index. When the Atkinson Review was published, activity data were available for residential care, nursing care, home care, day care, the provision of equipment and home adaptations, meal deliveries and referrals and care assessments undertaken. Residential, nursing and day care were further split by client group to reflect differences in the costs of providing care to different client groups.

However, this data collection was ended in 2013/14 and the collection that replaced it in 2014/15 covered a reduced set of activities, causing the proportion of ASC expenditure covered by the cost-weighted activity index to fall from 76 % to just 36 %. As a result, a new methodology has been developed (see Lewis (2018b)) which uses activity data, where available, to generate a cost-weighted activity index, and where it is not, calculates volume output using the 'outputs equals inputs' convention.

([20]) Local government in England is divided between 152 'top-tier' local authorities, with fewer in the smaller nations of Scotland and Wales. The Health and Social Care Service in Northern Ireland is a public body with responsibility for both health care and adult social care, although funding arrangements are similar to the rest of the United Kingdom, with health care being available free-at-the-point-of-use, but certain ASC services are means-tested and can charge clients directly.

While this is the best measure available for output, the loss of such a large proportion of activity data limits our ability to measure productivity across the whole ASC service sector, although separate productivity measures covering the service elements for which activity data remain (residential and nursing care) are produced to analyse these services specifically.

***Developing quality indicators for ASC services***

While activity data matching the requirements of the Atkinson Review were readily available in the 2000s, suitable quality measures were not.

The absence of available quality measures for ASC was not only a problem for the implementation of the Atkinson Review guidelines but also a problem for policy analysts trying to understand the performance of the ASC sector, whose main data source was the inspection reports of care homes carried out by the Care Quality Commission.

As a result, the ONS organised a cross-body programme, Measuring Outcomes for Public Service Users (MOPSU) to develop a toolkit for measuring ASC outcomes, along with other strands on building quality measures for early years education and measuring the third sector (see ONS (2010)). The MOPSU project on ASC outcomes was led by the Personal Social Services Research Unit (PSSRU) at the University of Kent providing sector-specific research and economic expertise.

At first glance, health care and adult social care may appear to be similar services, with both involving the care of individuals with health problems. However, while the main element of the health care quality adjustment measures the gain in health resulting from a hospital procedure performed at a point in time, the primary benefit from social care is an improvement in quality of life over the period social care is being received.

The project considered several approaches to measuring the outcomes of social care (see ONS (2007)), including:

- the extra-welfarist approach, where the desired outcome is pre-determined by the researcher and achievement against this outcome measured on a scale;
- the hedonic psychology approach, which involves studying clients' spontaneous approach/avoid, continue/desist and good/bad reactions at various moments in time as they use services;
- the capabilities and functioning approach, first developed by Sen (1985), which measures clients' opportunities or potential to obtain desirable 'functionings' such as being fed or having meaningful social relationships.

The approach taken followed the capabilities and functioning approach, and applied it to form a measure on a QALY-style zero to one scale, known as social care-related quality of life (SCRQoL). This is used as a quality adjustment on ASC output.

As described in the section on health care, there are several alternative questionnaire forms for measuring the quality of life element of QALYs, with the EQ-5D used in the NHS patient-reported outcome measures. The MOPSU project therefore needed to design a questionnaire for eliciting SCRQoL, based loosely on the capabilities described by Sen as essential elements of well-being.

An analysis of existing literature revealed eight broad domains which, with minimal overlap, appear to determine quality of life:

- personal cleanliness and comfort;
- accommodation cleanliness and comfort;
- safety;
- food and nutrition;
- control over daily life;
- occupation;
- social participation and involvement;
- dignity.

However, simply surveying care clients to rate their satisfaction on each of the eight domains against four possible responses for each creates two problems. Firstly, there is no reason to assume that each of the domains is of equal value to care clients — some may be more important to overall well-being than others. Secondly, it is not certain that the levels of responses that clients give against their experience (such as needs fully met, mainly met, partly met or not met) should be allocated a set of equally-spaced utility values, such as 1, 0.67, 0.33 and 0.

To deal with these issues, the Personal Social Services Research Unit (PSSRU) worked with RAND Europe on a study to determine the relative importance of each of the domains, and various 'levels' of experience within the domains, by asking care clients to rank the best and worst outcomes of a range of possible 'levels' of the above categories.

This study ([21]) enabled the construction of 'weights' for the preferences such that each 'level' of experience for each domain is attributed a utility value. Table 3 shows an example with two of the domains. The weights demonstrate that a difference in utility value between the top and bottom level responses for the control over daily life domain (the client having as much control over their daily life as they want and the client having no control over their daily life) is greater than the difference in utility value between the top and bottom responses for the social participation domain (the client having as much social contact as they want with people they like and the client having little social contact with people and feeling socially isolated). Of the eight domains, control over daily life had the greatest range in utility between the highest and lowest response, and this was bounded between zero and one. However, the utility weighting study also revealed that the difference between the first and second level response of each domain was lower for the control domain than for the social participation domain.

([21]) While the MOPSU project established the principles of weighting different domains of quality of life, the actual weights used in the quality adjustment are derived from a later study based on a number of specific surveys (see Netten et al. (2012)).

**Table 3:** Utility weights for two example domains

| Domain level | Utility weight |
|---|---|
| Control over daily life | |
| 1. I have as much control over my daily life as I want | 1.000 |
| 2. I have adequate control over my daily life | 0.919 |
| 3. I have some control over my daily life, but not enough | 0.541 |
| 4. I have no control over my daily life | 0.000 |
| Social participation and involvement | |
| 1. I have as much social contact as I want with people I like | 0.873 |
| 2. I have adequate social contact with people | 0.748 |
| 3. I have some social contact with people, but not enough | 0.497 |
| 4. I have little social contact with people and feel socially isolated | 0.241 |

*Source:* Netten et al. (2012)

## IMPLEMENTATION OF THE ADULT SOCIAL CARE QUALITY ADJUSTMENT

To collect the social care-related quality of life (SCRQoL) data needed to measure the performance of local authority social care services, the Adult Social Care Survey was introduced in April 2010 and now interviews over 10 % of adult social care clients in England annually. The measure of SCRQoL, along with other outcome measures from the Adult Social Care Survey, form the Adult Social Care Outcomes Framework, a set of indicators used to evaluate the performance of local authority ASC services across England.

While a change in the measure of SCRQoL gives a good indication of changes in the well-being of the care population, the measure does not give a definitive answer on whether a change in SCRQoL can be attributed to social care services or results from changes in the underlying care population or their wider environment. For instance, an improvement in the average response to the control over daily life question in Table 3 could result from improvements to the quality of care which result in clients being more involved in decisions about their care, but could also result from a change to the care population to include more lower-need clients whose health status may afford them more independence than other clients.

To produce an attributable quality adjustment, it is therefore necessary to develop a measure which isolates the effect of service quality on outcomes from the other factors which may also influence these outcomes. Adjusted social care-related quality of life (adjusted SCRQoL) was developed by the Quality and Outcomes of Person-Centred Care Research Unit (QORU) from the earlier work on SCRQoL to provide such a measure for the Adult Social Care Outcomes Framework (ASCOF) and was introduced into the 2016/17 indicator set.

The adjusted SCRQoL measure controls for a range of factors outside the control of social care providers which may affect SCRQoL including age, health status, the suitability of the clients' home for meeting their needs and the clients' ease of travelling around outside in their local environment through using regression analysis to derive an estimate for the expected effect of these factors on SCRQoL (Forder et al. (2016)).

While the adjusted SCRQOL measure has only been published in 2016/17 and for community care clients, the ASC output quality adjustment used by the ONS for community care is produced using the same parameters from data provided by the Adult Social Care Survey for the period 2010/11-2016/17.

For residential and nursing care, the quality adjustment is derived from a similar regression analysis informed by Yang, Forder and Nizalova (2017) and controls for:

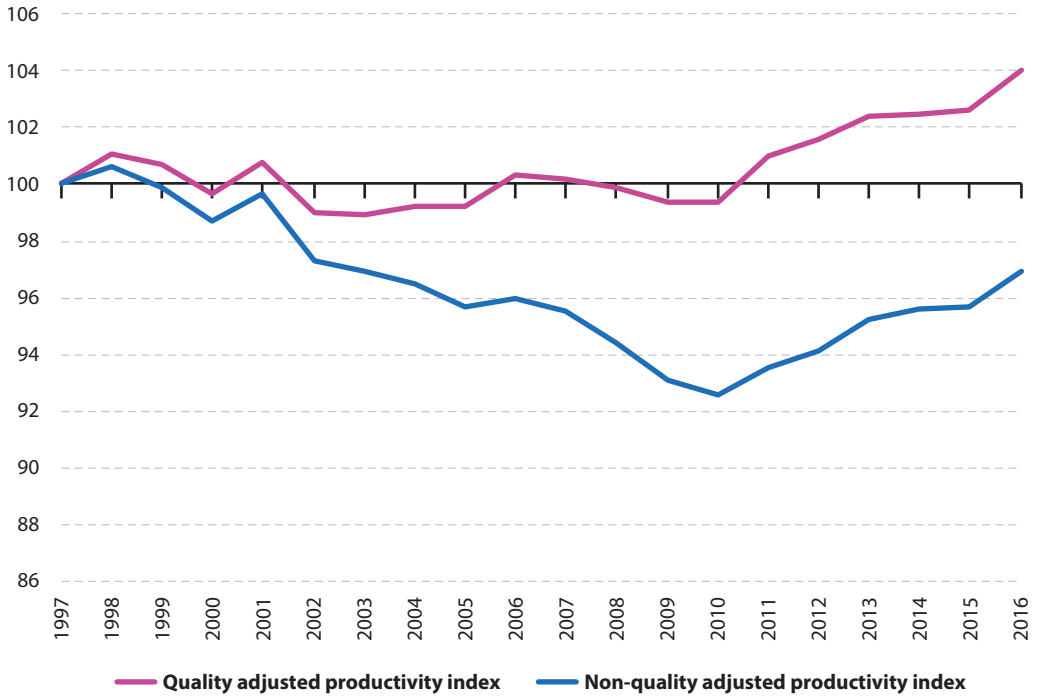
- gender;
- ethnicity;
- age;
- self-reported health status, level of pain and level of anxiety;
- the number of basic activities of daily living (ADLs) the client needs support with;
- whether the client can deal with their finances and paperwork.

The ASC quality adjustment therefore provides an estimate of the change in the key ASC outcome attributable to social care services.

## Aggregate impact

The aggregate impact of these quality adjustments on total public service productivity estimates are notable. In ONS (2019b) it was shown that non-quality adjusted public service productivity fell by 3.1 % between 1997 and 2016, while quality adjusted productivity rose by 4.0 %.

**Figure 3: Total public service productivity index, quality adjusted and non-quality adjusted, United Kingdom, 1997-2016**
(1997=100)



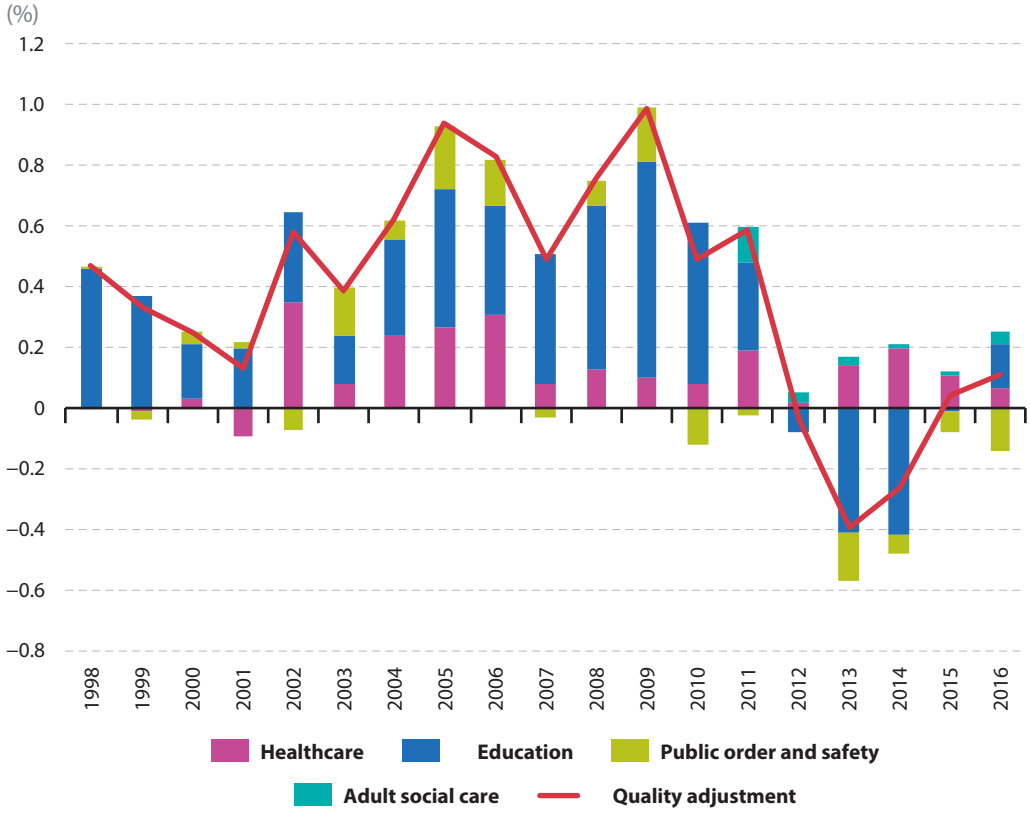

*Source:* Office for National Statistics — Public service productivity: total, United Kingdom, 2016 (ONS (2019b))

We can also break down this aggregate quality adjustment factor into contributions from the four components, as shown in Figure 4 where the education quality adjustment can be seen to be the largest contributor to overall public service quality, and thus that volatility in this series can induce substantial movements on the aggregate.

On average, the education quality adjustment has added 0.2 percentage points per year to growth in total quality adjusted output between 1998 and 2016. However, it acted as the main driver for the decline of overall quality between 2012 and 2015, averaging a negative contribution of 0.2 percentage points in this period, although we continue to explore method improvements here. In 2016, it returned to having a positive contribution of 0.1 percentage points.

For healthcare, the impact of its quality adjustment has been positive, with some variation in the size of its effect, contributing upwards in every year except 2001. The public order and safety quality adjustment, on the other hand, generally made upward contributions to the total rate up until 2010, but has since made consecutive downward contributions. This was due largely to the negative impact of the prison safety adjustment, reflecting increases in the number of self-harm and assault incidents reported in prisons. Finally, applied from 2011 onwards adult social care quality contributed positively.

**Figure 4: Contribution to total quality adjustment growth by service area, United Kingdom, 1998-2016**




Note: sum of components may not equal the total due to rounding. Healthcare quality adjustment applied from 2001 onwards. Adult social care quality adjustment applied from 2011 onwards.

*Source:* Office for National Statistics — Public service productivity: total, United Kingdom, 2016 (ONS (2019b))

# 4. The big issues in measuring public service outcomes

As outlined above, generating quality adjustments for a variety of public services is both feasible and capable of significantly improving the quality of the statistics being produced. In addition, where methodologies have been developed the authors believe there is a strong potential for other countries to use these as a substantive foundation upon which to build methodologies tailored to their countries service design.

However, in line with Atkinson's recommendations we do not believe it is feasible in the United Kingdom, or any country to stand still. Continuous improvement and development of these methods is required, and to make the most of this opportunity it is sensible to consider the key lessons which the current practices provide, particularly if other countries wish to learn from these examples.

Similarly, before discussing the welfare implications of outcomes which the public services contribute towards, this section addresses the key issues and 'lessons learnt' from a decade of attempting to apply the Atkinson principles in the United Kingdom context. These are:

- how should various aspects of quality change be valued and weighted?
- how should different quality adjusted services be weighted together?
- how do we keep pace with the rate of technological change?
- should we be following individuals or use aggregate data?
- what do we do when a change in policy affects our measure?
- where do we source objective weights?
- how do we trade-off consistency of estimates with different needs for data in relation to devolved matters?

## How should various aspects of quality change be valued and weighted?

There are two symmetric problems in relation to the valuation and weighting of quality change:

- what to do when a public service delivers multiple outcomes, which could all contribute towards the quality adjustment we calculate for that service; and
- what to do when a single outcome is impacted by multiple public services?

Clearly for the first of these, when a common metric exists which can be applied to multiple outcomes, such as quality-adjusted life years (QALYs) in health, this appears a trivial question: once the value of a single QALY is established healthcare interventions can be theoretically evaluated by comparing their cost to the value of the number of QALYs they deliver per course of treatment.

Complexities can, however, still emerge. As explained above, the QALY measure has two elements, a health-related quality of life element and a temporal element; and can therefore

combine the effect of improvements in health-related quality of life and increases in the length of life resulting from treatment. Data is required for both dimensions across the whole population to derive a quality adjustment, which is a non-trivial investment. Equally, while the increase in QALYs following an intervention would be far closer to a measure of the quality of health service provision than NHS performance indicators or broader outcome measures such as life expectancy, which would require attribution factors to be generated, the use of a quality adjustment solely based on QALYs would still face the problems described in Part 3.

So, given that in the United Kingdom we do not use QALYs, the health quality adjustment is applied to output as a simple scalar variable, such that a 1 % increase in the quality adjustment results in a 1 % increase in quality adjusted output.

But changes in quality may reflect changes in the value of the service that are less or greater than this simple scalar imposes. Deriving accurate valuations to either weight contributions to the quality adjustment, or to weight the quality adjustment vis-à-vis the outputs remains a formidable challenge, as illustrated by research by Ryan et al. (2014) on the case of valuing patient satisfaction. Whilst a range of methods are available, eliciting firm reliable values is at present almost impossible, and relies on an ability to calculate objective weights. This can make it difficult to diagnose the exact causes of quality change over time, a non-trivial complaint in a measure regularly used to inform public policy analysis.

The second scenario is perhaps most easily explained in relation to the way health and social care interact to support improved health outcomes, or the interactions between the various agencies within the criminal justice system (CJS). The CJS is a collection of agencies working in partnership towards a common goal. The effective functioning of the CJS requires the processing of offenders from arrest to prosecution, to the delivery of justice — whether punishment or acquittal. An accurate measure of the increment to collective welfare from the CJS should reflect this. This implies that one cannot treat the police, the courts, and the prisons as entirely individual, stand-alone entities: the effectiveness of each agent within the CJS depends, to varying degrees, on the effectiveness of the others. For example, the quality of prosecutions undertaken by the Crown Prosecution Service will depend on the quality of the investigative work undertaken by the police. In describing the CJS we have sometimes used the analogy of a car engine. Subsequently, the question becomes one of attributing a system-level outcome (reducing re-offending) across the various component parts of the CJS, when some of these, particularly prisons, have their specific quality measures (for example, safety and decency). This can lead to some parts of the system having a lower weight for reducing re-offending when they may be assumed to have a higher weight than others.

However, we continue to need to weight the activities which count towards output, and in doing so we need to weight their output by the quality measures, returning us to the first bullet point above. This brings together the different outcome measures when different parts of the overall system have different outcomes relating to them. In principle, it is most desirable to weight together different quality metrics or indicators based on the value placed by individuals and society on services and their various attributes. Such an approach was taken with the adult social care quality adjustment, using data from separate studies specifically commissioned to understand how social care clients valued different aspects of well-being. The stated preference approaches provided differential weights for different aspects of well-being and enabled different levels of responses to be assigned relative values.

Extending such an approach across all service sectors requires the conducting of a range of studies to identify such preferences. In each case, three key questions would need to be answered before extending this approach: whose preferences are used, how to reflect changes in preferences over time, and how to derive an 'average' valuation?

Another possibility may be to weight domains of quality according to the relative costs associated with them. This assumes that the revealed preference of service managers on what they consider important reflect social preferences. However, costs do not necessarily correlate with quality — this is where the Atkinson Review came into the story.

In the absence of adequate data then a default solution may be to tend towards equal weights. However, this method is clearly sub-optimal: the long-term robustness of equal weighting, lacking in empirical support, could undermine the validity of the associated measure produced. In relation to prisons, where re-offending, prison safety and preventing escapes come together as three clear outcomes we need to recognise, we used the approach of taking a weighted average of prison performance measures, as used by Her Majesty's Inspectorate of Prisons, on the basis that these were set by Ministers in Parliament, and as such could be argued to represent social preferences. This relies on assumptions of government efficacy in delivering this role but appeared justifiable over any other arbitrary set of weights. Such 'social preferences', however, cannot be observed in all areas of public services, although the experiment of using such performance structures may be replicable in other 'inspected services'.

## How should different quality adjusted services be weighted together?

The challenge described above in relation to a single service, become an even more complex matter when one begins to combine services together. The classic method used under Atkinson is to do this using cost weights. These are objective, and if one believes the marginal pound is efficiently allocated by government then should be reflecting equal value. However, if one considers the question: would a 10 % reduction in the quality of GBP 1 000 of spending on health be worth more or less than a 10 % reduction in the quality of GBP 1 000 of spending on forestry, one can immediately see that, by dint of the sheer differential in the volume of funding devoted to these two services, this would likely deliver very different marginal impacts and public reactions. Therefore, are cost weights appropriate if we cannot break down changes in costs between changes in the prices of output and changes in the quality of the outcomes delivered by these outputs? Diewert and Fox (2017) present arguments for alternative weighting approaches based upon the relative value to users, which the authors consider intuitively strong and worthy of significant further consideration.

## Keeping pace with new technology, systems and data

An additional recommendation raised by Atkinson was the need to maintain and continue to develop quality adjustments through time. Quality measures which were identified initially, particularly in periods of rapid technological change, may no longer be fit for purpose and

may, with the passing of time, fail to continue to measure the key underlying principle you are trying to measure. Principled measures are key, but must reflect change.

For example, the health quality adjustment itself remains largely unchanged since its introduction in 2005, but the range of metrics health policy analysts study has not. In 2010, a new set of indicators for measuring health care performance, the NHS Outcomes Framework (NHSOF), was introduced and has become the central source for analysts measuring health care outcomes. Having been produced prior to the NHSOF, the results of the current health care quality adjustment do not always triangulate with the story that health care policy analysts derive from the NHSOF. In 2016, the Centre for Health Economics at the University of York convened a workshop to bring together policy analysts and health economists to consider the criteria that should be used for selecting NHSOF indicators, and this was followed up with a paper, Bojke et al. (2018), applying the criteria to these indicators. The challenges of adopting NHSOF indicators within a quality adjustment exercise are considerable, as they are not drawn from a single data source, and so are published at different times and often variable frequencies. However, such a review demonstrates the need to regularly review quality measures to ensure their continued relevance as policies and data sources change and may lead to a quality adjustment with relevance to users.

The *quid pro quo* here is the allocation of development time by NSIs. In the authors' experience the trade-off between investing in updating existing quality adjustments versus the creation of new measures covering new service areas has regularly required consideration. In recent years, new service areas have been prioritised where developments in reporting and data sources have opened the door to creating a new adjustment at relatively low costs. The continued pace of change in health and education, the United Kingdom's two largest public services, however probably make the need to revert to revising their measures inevitable in the coming years.

In relation to new data, where it is difficult to forge a link through to an individual's experience, which is the approach followed in health and adult social care, we have found that the most practicable application is through the use of published data sources, which are generally aggregated at the population level to track the movement in group or average performance. Criminal justice is a prime example of this. Efforts to focus on individual offenders, or the 'offender journey' resulted in a failure to deliver a quality adjustment in this space until 2017, when the ONS changed tack to focus on aggregate performance data. The key to unlocking this was the delivery by the ONS of an experimental dataset on the relative severity of crime (capable of answering questions like 'how many burglaries equal a murder?'), which provided a set of objective weights to adjust raw re-offending data and provide a consistent measure of whether outcomes were improving or weakening through time.

This approach brought the additional benefit of allowing service providers to better engage with the productivity statistics, as they were grounded in concepts and measures they were currently working with and understood. Providing objective insights where these had been missing previously appears to the author's to be a positive direction of travel, particularly when this work could be delivered at little additional cost.

## What do we do when a change in policy affects our measure?

There are instances where the measure itself is subject to policy decisions, and is directly affected by policy change, not just in terms of the level, but also in terms of the definition of the measure itself. This is, broadly, always the case, but in some areas it is more pertinent than others. It is particularly the case when there are fears of 'gaming', that is where the definition of the measurement itself leads to undesired outcomes. Education is a prime example of where government policy has been shaped by the need to address a set of interlocking concerns.

Whilst the ONS has historically used the GCSE APS attainment as a quality metric, the actual application has changed noticeably over time, in response to three key issues, where there was a perceived threat that the measure had been corrupted ([22]). Firstly, there is the question of whether attainment through time has been consistently measured or suffered from 'grade inflation' ([23]), secondly have schools made greater use of 'easier' or more vocational courses to artificially inflate APS scores, and thirdly have schools improved marks by teaching to the test rather than giving a rounded education.

In light of this, the Department for Education established a review which found evidence of improvement in pupil's attainment in England over the period. However, when similar analysis was carried out on other measures and systems of pupil attainment used in the United Kingdom and within the OECD, they found, in contrast to the APS, little overall improvement in the level of pupil attainment. It is, however, worth noting that these findings were based on less timely data with much smaller sample sizes than national performance data, but they called into question the validity of GCSE APS data as a proxy either of educational attainment at that age, or as a proxy for the whole system, as the current quality adjustment implies.

To address these worries, reforms to GCSE grades were introduced by the Department for Education in 2014, following the Wolf Report (2011). This changed the qualifications eligible to count towards APS, particularly in relation to vocational qualifications on school performance measures in England ([24]). To reflect this an alternative approach to quality adjusting United Kingdom public service education output was proposed (ONS (2015a)) and adopted (ONS (2015b)).

The method replaced the use of APS data for England with Level 2 (or L2) attainment at age 16 for the years 2008 to 2013. Level 2 attainment equated to five or more GCSEs at grades A*-C or an equivalent (and eligible) Level 2 vocational qualification. This is a threshold measure of the percentage of students achieving a particular level of attainment, compared with the APS which takes into account the full distribution of attainment data, making Level 2 attainment less susceptible to changes in the education system and pupil behaviour. This is the current method used for quality adjustment in England. However, alongside these changes, in 2017 a further revision to GCSE grading was introduced which presented a more fundamental

([22]) Notwithstanding the fact that our experience of using a single measure (age 16 GCSE test results) as a proxy for performance across the age spectrum is that this model makes it difficult to reflect differential performance in one part of the education system (for example, primary or early years) against another (for example, secondary), when the quality measure simply does not capture more than one of these.

([23]) The converse argument is that, in the face of increasing tuition fees, and low wage growth in low skilled jobs, students have responded to market forces by investing more heavily in their own development whilst education is free, resulting in improving performance.

([24]) The significant increase in APS between 2008/2009 and 2011/2012 could partly be attributed to increases in the number of non-GCSE examinations taken because of changes in the type of examinations, which counted towards performance.

challenge. In a further effort to address perceptions of grade inflation, a new grading structure was introduced. This deliberately did not enable a one-to-one matching with the old banding structure, introducing computational challenges in preventing a discontinuity in the series.

**Table 4:** Old and new GCSE band equivalences

| Old structure | New structure |
|---|---|
| A* | 9 |
| | 8 |
| A | 7 |
| B | 6 |
| | 5 |
| C | 4 |
| D | 3 |
| E | 2 |
| F | 1 |
| G | |
| U | U |

Clearly, a measure which is the subject of frequent change is not a stable base upon which to build a long-term quality adjustment.

## How do we trade off consistency of estimates with different needs for data in relation to devolved matters?

As mentioned above in relation to the education quality adjustment, for reasons of data comparability and availability, the level of education quantity in primary and secondary schools in Northern Ireland is quality adjusted in line with that applied to English schools.

Similarly, while current measures and methodologies to reflect quality change in the CJS are applied to the output of the United Kingdom as a whole, the associated metrics reflect but a subset — covering England and Wales. Here the implicit assumption is made that changes in quality of the CJS in Scotland and Northern Ireland follow the trend observed in England and Wales.

Whilst only United Kingdom level estimates are produced we can, to some degree duck these issues, but in light of a growing need to provide statistics for devolved administrations and lower-level geographies it is clearly problematic to either attempt to compare an area whose quantity of output is quality adjusted with one which is not, or to quality adjust two areas by an adjustment factor derived in only one of them. In a world where decision-making powers in relation to these services have been devolved to administrations in each of the component countries of the United Kingdom it is clearly problematic for decision makers in Northern Ireland or Scotland to have to view the productivity of services they have responsibility for through a lens which can be argued to be distorting their view of their system relative to the other nations of the United Kingdom. This issue was explicitly recognised in Atkinson's Principle E.

This is exacerbated where different administrations or legal systems have resulted in long-standing differences between the model of services provided by the constituent countries, their methods of delivery and the machinery of government. These differences are set to become potentially more important because of devolution. Likewise, by applying common factors, we may well fail to reflect variations in priorities/desired outcomes, particularly as quality metrics and their associated weightings become more granular.

Given that many public services in the current model are not quality adjusted and at the aggregate level we are therefore regularly comparing quality adjusted and non-quality adjusted sectors, the ONS is exploring removing non-native quality adjustments ([25]) where these are currently applied.

# 5. The treatment of non-attributable outcomes as welfare gains

Atkinson and Parts 2-4 of this paper focus on the outcomes which are directly attributable to the activities and outputs of public services, however there is merit in stepping back to consider some fundamental questions about the exact scope under consideration and the implications of that scope on the object of interest: welfare gains.

There is a well-known difference between the evolution of outcomes which people value and the effect of public services in generating those outcomes, (see, for example, Stiglitz et al. (2009) or, Bean (2016)).

At its simplest, the Atkinson framework conceives that the volume of activity is not adequately measured by the outputs of that sector if insufficient attention is paid to quality change. For products in the market-sector this is captured through adjustments made to the deflator to decompose price changes into those caused by changes in the general price level and those caused by changes in the quality of the product. The relative price of a product should increase as its quality increases. This is obviously a more complex exercise when prices cannot be observed and public services, where such prices do not exist, exemplify this. The Atkinson Review therefore argued for the application of quality adjustments derived from directly observable data.

Are then these quality adjustments equivalent to the changes in consumers' welfare? The answer here from Atkinson is unequivocally 'no'. Increasing life expectancy, for example, is clearly of value but only a part of that can be attributed to improved health services. The majority of the rise is likely to belong to dietary and other lifestyle changes. So, there is an additional question as to how these wider effects can be measured. The Atkinson quality adjustments only capture that aspect of the welfare gain which is directly attributable to the public service.

([25]) In other words, applying English quality adjustments to Scottish or Northern Irish services.

However, in terms of the debates (summarised in Heys, Martin and Mkandawire (2019)) about measuring the modern economy and the need to understand why citizens increasingly view GDP as a poor proxy for welfare measures, this is a key point for two reasons:

- Whilst Stiglitz et al. (2009) encourage the focus to no longer be on improving GDP as a welfare measure, recent studies (Brynjolfsson et el. (2019), Hulten and Nakamura (2018)) show there remains an appetite for this approach because of the dominance of GDP within political debate. If public services are a significant fraction of GDP, and quality adjustments have a noticeable impact on volume growth in relation to these services, and this is not being taken into account, as it generally isn't, then this may introduce a wedge between GDP and welfare growth even if the concept of GDP, including public service quality adjustments, should share a common growth rate with welfare. This does not negate the thrust of arguments which suggest we should go 'beyond GDP' to measure welfare, but it remains valid to attempt to measure GDP growth as accurately as possible.
- Welfare gains from outcomes which relate to, but are not attributable to public services, may be a significant driver of any perceived difference in the behaviour of GDP and welfare, so if one wanted to identify a way to measure welfare, one would need to find a way to capture this element outside of GDP to contribute to a welfare measure.

This opens intriguing options:

- we know or can calculate the increased number of QALYs that a society is enjoying compared with some base year;
- we can also, using the methodologies set out and discussed earlier, place a value on each QALY reflecting the benefit society is estimated to receive from it;
- the simple product of the two gives an estimate of the increased welfare that society enjoys as a result of longer life expectancy or improved quality of life.

It should be emphasised that this is not a measure of public service output or would or should be used as a component of GDP. On the other hand, this measure of a key dimension of welfare is of importance and relevance in its own right. Further, a number of the data sources that would be needed are readily available — many of them, for example, are collected as part of the datasets being assembled for the sustainable development goals (SDGs). In turn, this should enhance the international comparability of such measures.

This suggests there are three topics of interest for future research:

- how to measure the welfare gains from increased life expectancy (for example Crafts (2002));
- how to measure the contribution to these welfare gains from public services, to improve public service output and productivity measurement; and,
- what is the relationship between the two?

To expand on the illustration described above, for example, the public service health quality adjustment is not denominated in pounds, but is a quantity uplift factor. Research is required to identify a method for getting both the wider welfare gain and the public service quality adjustment into the same base for valuation to allow comparison and evaluation.

# 6. Conclusions

This paper discusses possible ways forward in two related but different areas. It draws upon the United Kingdom's experience for this purpose.

One relates to the ongoing but unfinished agenda as to how to measure the outputs of goods and services which are 'free at the point of delivery', for the purposes of national accounts. Public services such as schools and health services are major examples of this kind. Over a decade ago, Sir Tony Atkinson provided a principled framework for this end. Consistent with the basic principles of national accounting, he advocated an approach by which this output should be measured as the value added by the services concerned. This value, in turn, equated to the improvement in outcomes directly attributable to the activities of the public services concerned.

Implementing this approach, as he recognised, is by no means straightforward, but the United Kingdom experience recounted above shows that strong progress can be made. Working with experts and practitioners, quantity and quality measures can be identified and used to give a good approximation of the value added by key public services, and thus their contribution to GDP. New data and intelligent use of existing data mean this can be done at low cost and in a way which maximises stakeholder understanding and acceptance.

But NSIs are also now grappling with a second task; measuring changes in welfare or more generally well-being, regardless of how they are generated. Health outcomes — for example, life expectancy or healthy life expectancy — are influenced by a variety of factors besides publicly-funded health services: diet, smoking prevalence and other lifestyle choices are obvious determinants. So, the central tasks under this agenda become first the identification of appropriate measures of outcome changes and then to determine how much value our societies place on those changes.

Adopting approaches based on clear principles, as Atkinson advocated, appears to be important for both agendas. For one thing, the outcomes used for the purposes of measuring the output of public services should be consistent with those used for measuring welfare more widely. Secondly, a principled approach helps to ensure intellectual rigour. Thirdly, international comparability is important. The specific circumstances and institutions of particular countries will vary and methodologies need to take this into account. Nevertheless, provided methodologies are all based on the same underlying principles, comparability can be safeguarded, particularly if they make use of commonly accepted and produced high level outcome measures.

Using widely recognised measures of well-being, such as life expectancy, enables us to create estimates of wider welfare measures to sit alongside GDP under the SNA. To answer our second question on the development of welfare measures, research is needed to understand the share of such gains attributable to public services.

For example, for education, the high-level outcome could be incremental additions to the stock of human capital (such as proposed by Jorgenson and Fraumeni (1992)). Improved human capital might be expected to lead not just to higher wages and salaries now but over a period of time, and hence consideration must be given to how the value of future expected wage returns should be discounted. To assess the value of public sector output in its contribution to human capital growth, and by extension, the productivity of publicly-funded education services, we would need to estimate the proportion of human capital growth attributable to publicly-funded education services. The additional growth in human capital beyond that created by education services would form a residual attributable to non-educational drivers of human capital in a welfare estimate or welfare account.

The challenges presented by these twin agendas are ones we believe the statistical community needs to take up. We are convinced that whilst implementation raises non-trivial issues, these are not insurmountable. If we chose not to do so, we would have little to say about the value of critical components of economic welfare, or the performance of a fifth or so of our respective economies. We would also miss a vital contributor to measuring the changing well-being of our societies. The cost of such a decision to our reputation would be profound. In a world where digital innovation is offering a stream of new free goods and services which undoubtedly add to welfare, missing flows of value such as those described above would cast any measure of welfare into doubt as incomplete and potentially misleading. The need to tackle these issues is both important and pressing. Failing to push on from the start that Atkinson established in this area would be a huge opportunity missed.

# Acknowledgements

The authors would like to thank Katherine Kent, Heather Bovill, Jonathan Athow, Richard Smith and two anonymous referees for their comments. With particular thanks to Josh Martin for his contributions towards this work. All errors remain the authors'.

# References

Atkinson, A. (2005), *Atkinson Review: Final report. Measurement of Government Output and Productivity for the National Accounts*, Palgrave Macmillan, Basingstoke.

Bean, Sir C. (2016), *Independent Review of Economic Statistics: Final Report, 2016*.

Bojke, C., A. Castelli, K. Grasic, A. Mason and A. Street (2018), 'Accounting for the quality of NHS output', *CHE Research Paper 153*, Centre for Health Economics, University of York, York.

Brynjolfsson, E., A. Collis, W. E. Diewert, F. Eggers, and K. J. Fox (2019), 'GDP-B: Accounting for the Value of New and Free Goods in the Digital Economy', *NBER Working Paper Series*, No. 25695, National Bureau of Economic Research, Cambridge, United States.

Castelli, A., M. Chalkley and I. R. Santana (2018), 'Productivity of the English National Health Service: 2015/16 Update', *CHE Research Paper 152*, Centre for Health Economics, University of York, York.

Crafts, N. (2002), 'UK Real National Income, 1950-1998: Some grounds for Optimism', National *Institute Economic Review*, Volume 181, Issue 1, pp. 87-95.

Dawson, D., H. Gravelle, M. O'Mahony, A. Street, M. Weale, A. Castelli, R. Jacobs, P. Kind, P. Loveridge, S. Martin, P. Stevens and L. Stokes (2005), 'Developing new approaches to measuring NHS outputs and productivity', National Institute of Economic and Social Research, *NIESR Discussion paper No. 264/CHE Research Paper 6*, Centre for Health Economics, University of York, York.

Diewert, W. E. (2011), 'Measuring productivity in the public sector: some conceptual problems', *Journal of Productivity Analysis*, Volume 36, Issue 2, pp. 177-191.

Diewert, W. E. and K. J. Fox (2017), 'Productivity Measurement in the Public Sector: Theory and Practice', Vancouver School of Economics, University of British Columbia, Vancouver.

Ford, G. and J. Lewis (2018), 'UK Health Accounts: 2016', Office for National Statistics, United Kingdom.

Forder, J., A. M. Towers, J. Caiels, J. Beadle-Brown and A. Netten (2008), 'Measuring Outcomes in Social Care: Second Interim Report', Personal Social Services Research Unit, University of Kent, Canterbury.

Forder, J., J. Malley, S. Rand, F. Vadean, K. Jones and A. Netten (2016), 'Identifying the impact of adult social care: Interpreting outcome data for use in the Adult Social Care Outcomes Framework', Quality and outcomes of person-centred care policy research unit (QORU), University of Kent, Canterbury.

Foxton, F. (2018a), 'Public service productivity estimates: total public service, UK: 2015', Office for National Statistics, United Kingdom.

Foxton, F. (2018b), 'Quality adjustment of public service public order and safety output: current method', Office for National Statistics, United Kingdom.

Glover, D. and J. Henderson (2010), 'Quantifying health impacts of government policies', Department of Health, United Kingdom.

Heys, R., J. Martin, and W. Mkandawire (2019), 'GDP and Welfare: A spectrum of opportunity', *ESCoE Discussion Paper No. 2019-16*, Economic Statistics Centre of Excellence, National Institute of Economic and Social Research, London.

Hicks, Sir J. R. (1941), 'The Rehabilitation of Consumers' Surplus', *The Review of Economic Studies*, Oxford University Press, Volume 8, Issue 2, pp. 108-116.

Hulten, C. and L. Nakamura (2018), 'Accounting for Growth in the Age of the Internet: The Importance of Output-Saving Technical Change', *NBER Working Paper Series*, No. 23315, National Bureau of Economic Research, Cambridge, United States.

Jorgenson, D. and B. Fraumeni (1992), 'The Output of the Education Sector' in *Output Measurement in the Service Sectors*, Z. Griliches, ed., National Bureau of Economic Research, University of Chicago Press, pp. 303-341.

The King's Fund (2015), 'Inequalities in life expectancy — Changes over time and implications for policy', August 2015.

The King's Fund (2017), 'What's going on with A&E waiting times?', retrieved 2 August 2018.

Kuznets, S. (1937), *National Income and Capital Formation, 1919-1935*, National Bureau of Economic Research, New York.

Kuznets, S., L. Epstein and E. Jenks (1941), *National Income and Its Composition, 1919-1938, Vol. 1*, National Bureau of Economic Research, New York.

Lewis, J. (2018a), 'Public service productivity estimates, healthcare: 2015', Office for National Statistics, United Kingdom.

Lewis, J. (2018b), 'Measuring adult social care productivity in the UK and England: 2016', Office for National Statistics, United Kingdom.

Mackillop, E. and S. Sheard (2018), 'Quantifying life: Understanding the history of Quality-Adjusted Life-Years (QALYs)', *Social Science & Medicine*, Volume 211, August 2018, pp. 359-366.

Mason, A., P. Ward and A. Street (2011), 'England: The Healthcare Resource Group system' in *Diagnosis-Related Groups in Europe*, R. Busse et al., ed., European Observatory on Health Systems and Policies Series, Open University Press, Maidenhead, pp 197-220.

McGinnis, J. M., P. Williams-Russo and J. Knickman (2002), 'The case for more active policy attention to health promotion', *Health Affairs*, Volume 21, Number 2, pp. 78-93.

Ministry of Justice (2016), 'Prison Safety and Reform', Ministry of Justice, United Kingdom.

Netten, A., P. Burge, J. Malley, D. Potoglou, A. M. Towers, J. Brazier, T. Flynn, J. Forder and B. Wall (2012), 'Outcomes of social care for adults: developing a preference-weighted measure', *Health Technology Assessment*, Volume 16, No. 16, . pp. 1-166.

NHS Digital (2018), 'NHS Outcomes Framework Indicators — August 2018 Release', retrieved 26 October 2018.

ONS (2007), 'Initial report: quality measurement framework project', Office for National Statistics, United Kingdom.

ONS (2010), 'Measuring Outcomes for Public Service Users', Office for National Statistics, United Kingdom.

ONS (2014), 'Health expectancies at Birth and at Age 65 in the United Kingdom: 2009-11', Office for National Statistics, United Kingdom.

ONS (2015a), 'Methods changes in Public Service Productivity Estimates: Education 2013', Office for National Statistics, United Kingdom.

ONS (2015b), 'Public Service Productivity Estimates: Education 2013', Office for National Statistics, United Kingdom.

ONS (2016), 'Research outputs: developing a Crime Severity Score for England and Wales using data on crimes recorded by the police', Office for National Statistics, United Kingdom.

ONS (2017a), 'Health state life Expectancies, UK: 2014 to 2016', Office for National Statistics, United Kingdom.

ONS (2017b), 'National life tables, UK: 2014 to 2016', Sanders, S., Office for National Statistics, United Kingdom.

ONS (2018), 'Estimates of the population for the UK, England and Wales, Scotland and Northern Ireland', Office for National Statistics, United Kingdom.

ONS (2019a), 'A guide to quality adjustment in public service productivity measures', Harris, L., Office for National Statistics, United Kingdom.

ONS (2019b), 'Public service productivity: total, UK, 2016', Campbell, S. and F. Foxton, Office for National Statistics, United Kingdom.

Ryan, M., P. Kinghorn, V. A. Entwistle and J. J. Francis (2014), 'Valuing patients' experiences of healthcare processes: towards broader applications of existing methods', *Social Science & Medicine*, Volume 106, April 2014, pp 194-203.

Ryen, L. and M. Svensson (2014), 'The Willingness to Pay for a Quality Adjusted Life Year: A Review of the Empirical Literature', *Health Economics*, Volume 24, Issue 10, pp. 1 289-1 301.

Sen, A. (1985), *Commodities and Capabilities*, Elsevier Science & Technology, Amsterdam.

Schreyer, P. (2012), 'Output, Outcome and Quality Adjustment in Measuring Health and Education Services', *Review of Income and Wealth*, Series 58, No. 2, pp. 257-278.

Stiglitz, J. E., A. Sen and J-P. Fitoussi (2009), 'Report by the Commission on the Measurement of Economic Performance and Social Progress'.

van Loon, M. S., K. M. van Leeuwen, R. W. J. G. Ostelo, J. E. Bosmans and G. A. M. Widdershoven (2018), 'Quality of life in a broader perspective: Does ASCOT reflect the capability approach?', *Quality of Life Research*, Volume 27, Issue 5, pp. 1 181-1 189.

Wolf, A. (2011), 'Review of Vocational Education — The Wolf Report'.

Yang, W., J. Forder and O. Nizalova (2017), 'Measuring the productivity of residential long-term care in England: methods for quality adjustment and regional comparison', *The European Journal of Health Economics*, Volume 18, Issue 5, pp. 635-647.

# Annex A: The Atkinson Principles

As drawn from pp. 55-56 of Atkinson (2005).

**Principle A**: the measurement of government non-market output should, as far as possible, follow a procedure parallel to that adopted in national accounts for market output.

**Principle B**: the output of the government sector should in principle be measured in a way that is adjusted for quality, taking account of the attributable incremental contribution of the service to the outcome.

**Principle C**: account should be taken of the complementarity between public and private output, allowing for the increased real value of public services in an economy with rising real GDP.

**Principle D**: formal criteria should be set in place for the extension of direct output measurement to new functions of government. Specifically, the conditions for introducing a new directly measured output indicator should be that (i) it covers adequately the full range of services for that functional area, (ii) it makes appropriate allowance for quality change, (iii) the effects of its introduction have been tested service by service, (iv) the context in which it will be published has been fully assessed, in particular the implied productivity estimate, and (v) there should be provision for regular statistical review.

**Principle E**: measures should cover the whole of the United Kingdom; where systems for public service delivery and/or data collection differ across the different countries of the United Kingdom, it is necessary to reflect this variation in the choice of indicators.

**Principle F**: the measurement of inputs should be as comprehensive as possible and in particular should include capital services; labour inputs should be compiled using both direct and indirect methods, compared and reconciled.

**Principle G**: criteria should be established for the quality of pay and price deflators to be applied to the input spending series; they should be sufficiently disaggregated to take account of changes in the mix of inputs and should reflect full and actual costs.

**Principle H**: independent corroborative evidence should be sought on government productivity, as part of a process of 'triangulation', recognising the limitations in reducing productivity to a single number.

**Principle I**: explicit reference should be made to the margins of error surrounding national accounts estimates.

# Annex B: Estimating public service quantity output

The process is carried out in several steps:

1. Time series data are compiled examining (a) the number of differentiated activities and (b) the level of expenditure in each individual sector, at the available geographic granularity.
2. A chain-linked Laspeyres volume index of output is produced for each educational sector such that:

$$\psi_t = \psi_{t-1}\left(\sum_i\left(\frac{\left(\left(a_{i.j,k,t}\right)-\left(a_{i.j,k,t-1}\right)\right)}{a_{i.j,k,t-1}} * \frac{x_{i,j,k,t-1}}{\sum_j x_{i,j,k,t-1}}\right)+1\right)$$

Where:
*i, j, k* and *t* index individual sectors, differentiated activities, geographical area and time respectively
$\psi_t$ is a chain-linked Laspeyres index of quantity output
$a_t$ is the number of activities
$x_t$ is the level of expenditure in current price terms

Output in the initial period (*t*=0) is set equal to 100.

3. A United Kingdom-level, chain-linked Laspeyres volume index of output is calculated using the individual sector indices and the relative cost weights, such that:

$$\Psi_t = \Psi_{t-1}\left(\sum_i\left(\frac{\psi_{i,t}-\psi_{i,t-1}}{\psi_{i,t-1}} * \frac{x_{i,t-1}}{\sum_i x_{i,t-1}}\right)+1\right)$$

Where:
*i* and *t* index individual sectors and time respectively
$\Psi_t$ is a chain-linked, aggregate United Kingdom, Laspeyres index of quantity output
$\psi_t$ is a chain-linked Laspeyres index of individual sector quantity output
$x_t$ is the level of expenditure in current price terms.

Output in the initial period (*t*=0) is set equal to 100.

The result of this process is a chain-linked, United Kingdom-level, Laspeyres index of quantity output for the respective service area. There are several equivalent methods of generating this result. In particular, this approach is equivalent to first calculating the indices for geographical areas and then aggregating over educational sectors.

# Annex C: Estimating public service quality adjusted output

The process is carried out in several steps:

1. The quality adjustment measures are converted into indices such that:

$$q_{i,j,k,z,t} = q_{i,t-1}\left(\frac{\beta_{i,j,k,z,t} - \beta_{i,j,k,z,t-1}}{\beta_{i,j,k,z,t-1}}\right)$$

   Where:
   *i*, *j*, *k*, *z* and *t* index individual sectors, differentiated activities, geographical area, quality measures and time respectively
   $\beta_t$ is respective quality metric
   $q_t$ is the level of quality achieved in delivery
   $q_{i,t=0}$ equals 1.

2. A chain-linked Laspeyres volume index of quality adjusted output is produced for each individual sector such that:

$$I_t^Q = I_{t-1}^Q\left(\sum_i\left(\frac{\left(\left(a_{i.j,t}q_{i,t}\right)-\left(a_{i.j,t-1}q_{i,t-1}\right)\right)}{a_{i.j,t-1}q_{i,t-1}} * \frac{x_{i,j,t-1}}{\sum_j x_{i,j,t-1}}\right)+1\right)$$

   Where:
   *i*, *j* and *t* index educational sectors, geographical area and time respectively
   $I_t^Q$ is a chain-linked Laspeyres index of quality adjusted output
   $a_t$ is the number of activities
   $q_t$ is the level of quality achieved in delivery
   *xt* is the level of expenditure in current price terms

   Output in the initial period (*t*=0) is set equal to 100.

   For sectors which are not explicitly quality adjusted, $q_{i,t} = q_{i,t-1} = q_{i,t=0} = 1$.

3. As before, a United Kingdom-level, chain-linked Laspeyres volume index of quality adjusted output is calculated using the individual sector indices and the relative cost weights, such that:

$$L_t^Q = L_{t-1}^Q\left(\sum_i\left(\frac{I_{i,t}^Q - I_{i,t-1}^Q}{I_{i,t-1}^Q} * \frac{x_{i,t-1}}{\sum_i x_{i,t-1}}\right)+1\right)$$

   Where:
   *i* and *t* index educational sectors and time respectively
   $L_t^Q$ is a chain-linked, aggregate United Kingdom, Laspeyres index of quality adjusted output
   $I_t^Q$ is a chain-linked Laspeyres index of quality adjusted output for each individual sector.

# 2 Extended supply and use tables for Belgium: where do we stand?

BERNHARD MICHEL [1][2], CAROLINE HAMBŸE [1], BART HERTVELDT [1] AND GUY TRACHEZ [1]

**Abstract**: The construction of extended supply and use and input-output tables has been presented as a means of addressing some of the current challenges for the national accounts that are induced by economic globalisation. Such tables take into account within-industry firm heterogeneity that is not related to product characteristics, through a disaggregation of industries by size, ownership, exporter status or other relevant criteria. Beyond their contribution to improving national accounts data, these extended tables also allow us to derive new results on the participation of categories of firms in domestic and global value chains. In this article, we give an overview of where we stand today in terms of the construction of extended supply and use and input-output tables (SUT and IOT) for Belgium: methodological choices, data used, tables that have already been constructed, and analytical results. This is designed as an input for organising future work on extended tables for Belgium, but it may also provide useful information for other countries that engage in the construction of extended SUT and IOT.

**JEL codes**: C67, C81, D57, F14, F23

**Keywords**: economic globalisation, extended supply and use tables, input-output tables, firm heterogeneity, exporters, ownership, firm size

(1) Federal Planning Bureau, Belgium.
(2) Department of Economics, Ghent University.

# 1. Introduction

Globalisation brings new challenges for the national accounts and related economic statistics. One of these pertains to the industry breakdown in supply and use and input-output tables. Traditionally, industries in these tables group together producers according to the type of goods and services they produce. Within these industries defined in terms of product similarity, technological homogeneity ([4]) has been taken for granted. However, as value chains have become increasingly fragmented and global, within-industry patterns of specialisation have developed, which do not depend on the types of products delivered but are related to other characteristics of producers such as size, ownership or exporter status. This idea is in line with prior empirical research on firm heterogeneity (for example Bernard et al. (2009)). The aim of so-called extended supply and use and input-output tables (SUT and IOT) is to take such heterogeneity into account, in other words, to construct tables in which industries are disaggregated according to these characteristics.

Recently, there has been growing interest for such extended tables and the OECD and Eurostat have encouraged national statistical offices (NSOs) to start producing them. Their construction may serve various objectives beyond obtaining an industry breakdown in the SUT and IOT with greater within-industry homogeneity in terms of input structure. From a statistical point of view, work on extended SUT may contribute to improving the construction process of the regular annual SUT, for example in terms of the balancing process (OECD (2015)). In turn, this may lead to an improvement of the underlying national accounts. Furthermore, extended SUT with a breakdown by ownership provide an integrated framework for separating out the activities of multinational enterprises, which is advocated as an important step in addressing the challenges of recent developments in globalisation for the national accounts (see Ahmad (2018), Moulton and van de Ven (2018)). From an analytical point of view, fully-fledged extended IOT enable an enhanced value chain analysis. They allow us to correct for the downward bias in the import content of exports that is due to averaging import intensities over different types of producers within industries (Piacentini and Fortanier (2015)). They also make it possible to identify how different types of firms integrate into domestic and global value chains (Michel et al. (2018)). Further issues that could be addressed in the framework of national or global extended IOT are, for example, the distribution of income in value chains (Ahmad (2018)) or the bias in estimates of value chain employment (Miroudot (2016)).

For Belgium, work on extended SUT and IOT was launched at the Federal Planning Bureau (FPB) in 2017 (HeterIO project). It was decided to start with a disaggregation of manufacturing industries by exporter status in the 2010 tables as a test case to determine the feasibility of producing extended SUT and IOT for Belgium. The approach was to disaggregate industries in the existing conventional SUT based on the most detailed firm-level data that was used in the construction of these conventional tables. By using all available firm-level data for industry disaggregation in each stage of the production process of extended SUT, we strived to limit as much as possible the use of proportionality assumptions for disaggregations, in particular for input structures. This test case has yielded interesting results from both a statistical and an analytical point of view (Michel et al. (2018)) and raised support for further work on extended SUT and IOT for 2015.

([4]) What we refer to here as technology corresponds to input cost structures in the context of monetary supply and use and input-output tables.

The aim of this contribution is to provide an update on the work on extended SUT and IOT for Belgium: to briefly cover what has been achieved so far in terms of data, methodology and analytical results, and to give a structured overview of outstanding issues. This is designed as an input for organising future work on extended SUT and IOT for Belgium, but it may also provide useful information for other countries that engage into the construction of extended SUT and IOT.

# 2. Statistical work

The HeterIO project for constructing extended SUT and IOT for Belgium was launched at the FPB in 2017 with the disaggregation of manufacturing industries in the 2010 SUT according to exporter status. Exporter status was a straightforward choice because trade data are readily available and already used in the process of constructing the conventional SUT and IOT, and because Belgium is a small and very open economy for which this disaggregation criterion is of particular interest. The disaggregation was done for 2010 ([5]) since this was then still the most recent IO reference year and it was restricted to manufacturing industries ([6]) to keep the workload manageable ([7]). Industries were disaggregated at the most detailed breakdown available for the Belgian SUT. For this purpose, we used the full set of individual firm-level data that serve for the construction of the country's conventional SUT and IOT. This construction is based on data for legal units, which we refer to as firms. All these firm-level data share a unique identifier for each firm.

In practice, we proceeded in several steps, which are summarised in general terms in Figure 1.

a) We calculated the share of exports in turnover for all 40 194 manufacturing firms in the 2010 business register for the Belgian national accounts based on firm-level export and turnover data, and we defined exporter status as follows: a firm is considered export-oriented if exports represent 25 % or more of its turnover. All other manufacturing firms are considered as domestic market firms, in other words, firms mainly serving the domestic market. Overall, there were 2 430 export-oriented manufacturing firms in 2010 (6 % of all manufacturing firms). They accounted for 75 % of manufacturing turnover and 97 % of manufacturing exports ([8]). For defining exporter status, we decided to apply this relative threshold rather than separating out all exporting firms because we believe that it enhances homogeneity in terms of input structures within the resulting groups of firms. As such, we consider that the input structures of 'small exporters', in other words, firms that export less than the threshold share of their turnover, are more like the input structures of firms that do not export. This is in line with work on extended SUT for Denmark (Nilsson et al. (2019)). The choice of the threshold level (25 %) is, of course, arbitrary. Nonetheless, it has allowed us to avoid for most industries that the sample size gets too small for either export-oriented or domestic market firms (see also Point c)) ([9]).

([5]) The underlying conventional table is the 2010 SUT established according to the rules of the 2010 European System of Accounts (ESA 2010; see FPB (2015)).
([6]) NACE Rev. 2 10-33 broken down into 57 individual industries; NACE stands for the statistical classification of economic activities in the European Community.
([7]) The choice of manufacturing industries was motivated by the importance of manufacturing exporters for the integration into global value chains (see Michel et al. (2018)).
([8]) Since domestic market firms may export up to 25 % of their turnover, this group of firms actually accounts for a small share of total exports. In our data for 2010, this share amounted to 3 %.
([9]) It is our aim to test alternative threshold percentages in future work.

b) We disaggregated industry-level totals for output and intermediate input purchases (from the national accounts, column totals in the conventional SUT) based on the shares of export-oriented manufacturers in respectively total industry-level turnover and total industry-level purchases ([10]). Disaggregated value added (including net taxes on products) was obtained by difference.

c) We estimated the product distribution of output and intermediate inputs (columns in the SUT) for export-oriented and domestic market firms in manufacturing industries based on a restricted sample of firms for which we have information on turnover and purchases by product category. This information comes from extra questionnaires annexed every five years — in I-O reference years — to the structural business statistics (SBS) survey for mainly big firms ([11]). It covered 1 710 manufacturing firms in 2010 of which 980 were export-oriented ([12]). The advantage of this data situation is that we were able to perform a data-driven disaggregation of the columns of the supply and use tables for most manufacturing industries ([13]). We applied a RAS ([14]) procedure to ensure consistency with respect to the product distribution of output and intermediate inputs of the corresponding manufacturing industries in the conventional SUT ([15]).

**Table 1: Heterogeneous input-output table for Belgium, 2010**
(EUR million)

| | Export-oriented manufacturers | Domestic market manufacturers | Other industries | Domestic final demand | Exports of goods | Exports of services | Total output |
|---|---|---|---|---|---|---|---|
| **Export-oriented manufacturers** | 15 335 | 3 866 | 11 482 | 12 446 | 101 566 | 4 609 | 149 304 |
| **Domestic market manufacturers** | 6 900 | 5 697 | 14 730 | 13 278 | 8 975 | 2 888 | 52 467 |
| **Other industries** | 28 279 | 13 379 | 170 886 | 258 311 | 18 180 | 60 303 | 549 337 |
| **Imports** | | | | | | | |
| Manufacturing | 39 416 | 9 839 | 15 879 | 35 285 | 61 374 | 0 | 161 793 |
| Other | 26 526 | 3 558 | 49 175 | 7 382 | 14 312 | 0 | 100 952 |
| **Value added** | 32 848 | 16 128 | 287 186 | | | | |
| **Total output** | 149 304 | 52 467 | 549 337 | | | | |

([10]) Data on turnover and purchases is drawn from one of the following three sources: firms' annual accounts, their answers to the structural business statistics survey, and their annual VAT declaration.
([11]) As for all other firm-level data that we have used, these extra SBS questionnaires on the product detail of turnover and purchases are also used in the construction of conventional SUT for Belgium.
([12]) These 1 710 firms accounted for more than 78 % of total turnover in Belgian manufacturing in 2010.
([13]) Due to insufficient sample sizes for one of the two groups of firms, the output and input columns of 9 out of 57 industries had to be disaggregated proportionally based on shares of the two groups of firms in total turnover and purchases. These industries accounted for 23 % of output and 8 % of value added.
([14]) RAS is a bi-proportional scaling method. For a good overview of this method, see pp. 480-487 in United Nations (2018).
([15]) In the absence of detailed data, we disaggregated manufacturing industries in valuation tables proportionally for the transformation of uses to basic prices.

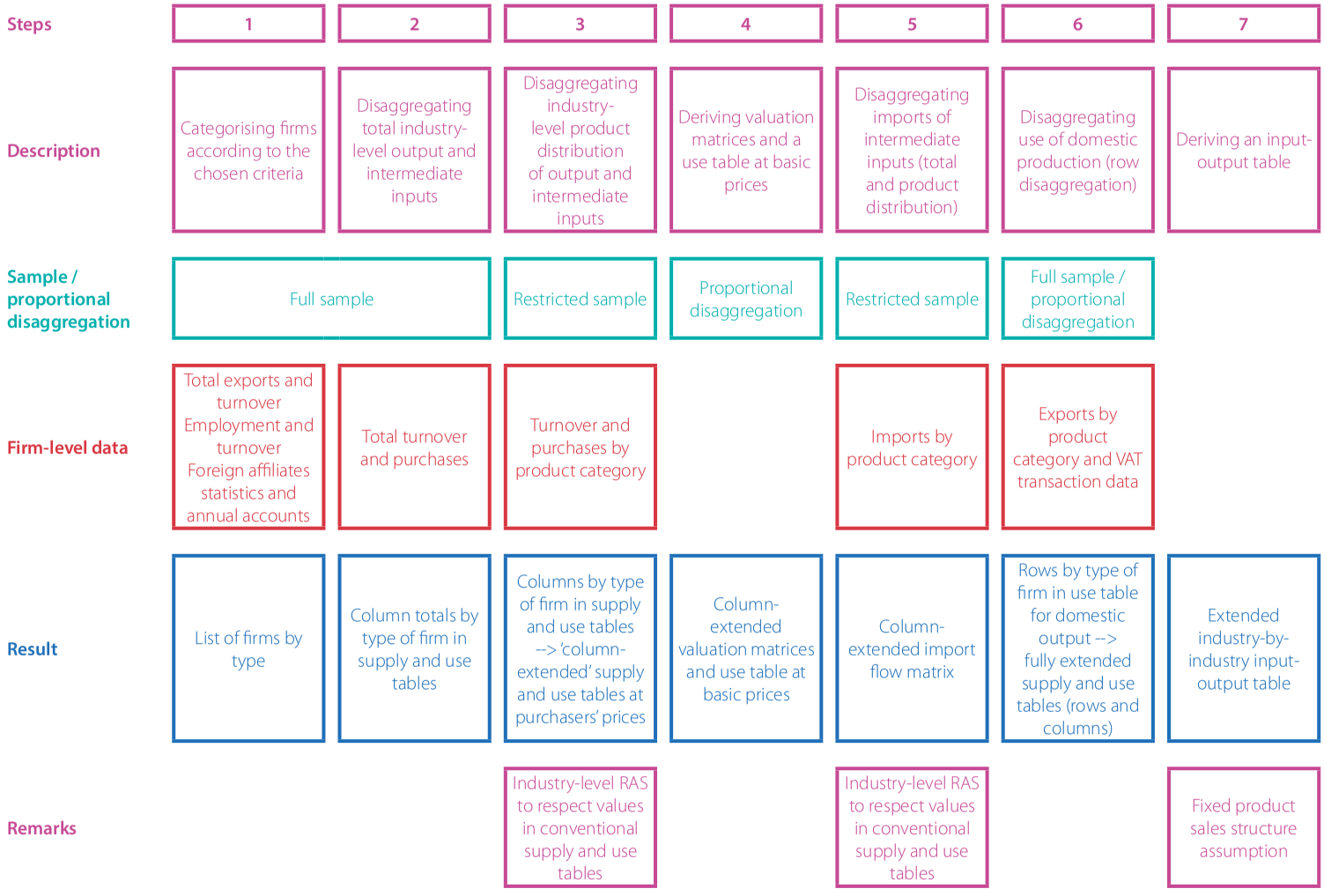
d) We disaggregated the columns of the import flow matrix (use of imported intermediate inputs) into export-oriented and domestic market firms in manufacturing industries based on product-level import data for both types of firms. Again, we applied a RAS procedure to make sure that the results respected the import flow matrix of the conventional use table. The use of domestically-produced intermediate inputs by export-oriented and domestic market manufacturers was calculated by difference. The further disaggregation of the use of domestic output by producing type of firm (export-oriented or domestic market manufacturers) was done proportionally (row split).

e) We derived an industry-by-industry extended IOT from the extended SUT in basic prices based on the fixed product sales structure assumption (Eurostat (2008)). The results are shown in very aggregated form in Table 1. As a last step, we integrated the extended IOT for Belgium into the 2010 global IOT from the 2016 release of the World Input-Output Database (WIOD) ([16]). For this purpose, we disaggregated Belgium's exports and imports by product category and partner country for all industries in the extended IOT, in other words, for manufacturing industries disaggregated into export-oriented and domestic market firms.

f) We are currently working on a disaggregation of the compensation of employees between export-oriented and domestic market manufacturers for a first look at the distribution of income. This can be done based on firm-level wage cost data. But it remains to be investigated whether and how other components of value added (taxes less subsidies on production, consumption of fixed capital, net operating surplus) can be disaggregated at the industry level. Furthermore, we have started to work on a disaggregation of industry-level employment. For total industry-level employment, this is based on administrative data from social security records. For employment by educational attainment, this is based on data from the so-called social balance sheet, which is an extra section of firm's annual accounts that contains information on employment.

([16]) For a detailed description of the WIOD project and the sources and methodology for constructing the global IOT, see Timmer et al. (2015).

**Figure 1:** Stages in the production process of extended SUT



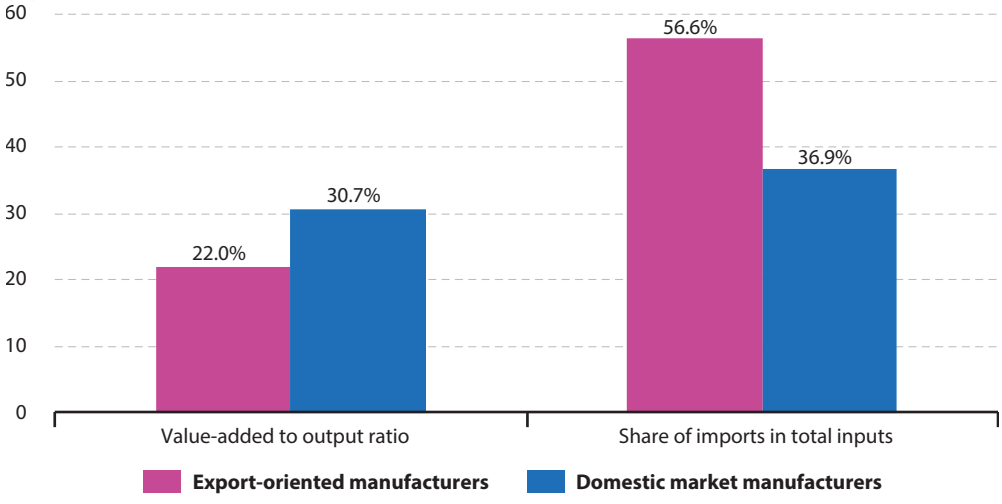
# 3. Results

From this extended IOT for Belgium for 2010, we have derived several results in terms of direct cost and input structures and in terms of value chain integration.

On average, export-oriented manufacturers have a lower value added to output ratio than domestic market manufacturers, in other words, they purchase proportionally more intermediate inputs (see left-hand side of Figure 2). We have used an independent t-test to determine whether the difference in value added to output ratios between export-oriented and domestic market firms is significant for our sample of manufacturing industries ([17]). We obtain a value of -2.74 for the test statistic, which yields a p-value < 0.01 for the two-tailed test, in other words, the difference in value added to output ratios between export-oriented and domestic market manufacturers is statistically significant ([18]).

Export-oriented manufacturers import proportionally more of the intermediate inputs they use (see right-hand side of Figure 2), in other words, export-oriented manufacturing firms engage more in offshoring, which reflects the greater cross-border fragmentation of their production processes. These results confirm prior findings on differences in import propensities between exporters and non-exporters based on firm-level data (for example Eaton et al. (2004); Bernard et al. (2009)). As for value added to output ratios, we find that the differences in import shares of these two groups within our sample of manufacturing industries are statistically significant (test-statistic of -3.29 and p-value < 0.01) ([19]).

**Figure 2: Direct production cost structures in manufacturing, Belgium, 2010**

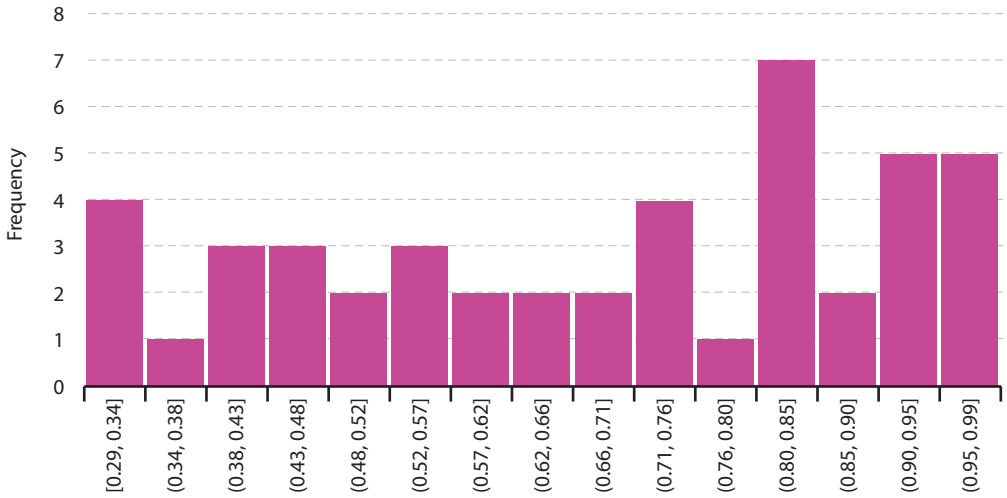



([17]) Prior contributions have used the Wilcoxon signed-rank test instead (for example Chong et al. (2019); Nilsson et al. (2019)). In our case, the standard independent t-test seems more appropriate given that our data fulfils the underlying assumptions (approximately normally distributed data within each group and equal variances across groups). Moreover, the non-parametric Wilcoxon test is used for comparing the medians of a sample before and after a treatment to check whether the treatment has a significant effect. Hence, it implies considering the criterion for disaggregation — exporter status in our case — as the treatment. This is questionable because it does not reflect a change in status.

([18]) See also the scatterplot of value added to output ratios in the *Data appendix*.

([19]) Since the variance for this indicator differs between the two groups, we have used Welch's t-test, which is more accurate than the standard t-test in case of inequality of the variances. The scatterplot for the import shares is also provided in the *Data appendix*.

**Figure 3: Distribution of the industry-level correlations between technical coefficients of export-oriented and domestic market manufacturers**




The product distribution of intermediate inputs differs between export-oriented and domestic market manufacturers. This is revealed by the histogram in Figure 3, which shows the distribution of the correlation between the technical coefficients of export-oriented firms and those of domestic market firms across all manufacturing industries ([20]); it excludes proportionally disaggregated industries. The distribution of these correlation coefficients is not skewed towards a value of 1, and the average correlation between the intermediate input structures of export-oriented and domestic market firms in the same industry is 0.71. This shows that export-oriented and domestic market manufacturers differ not only in terms of their propensity to purchase and import the inputs they use but also in terms of the types of inputs they use. In other words, we find evidence of technological differences between these two types of manufacturing firms, which is in line with the initial hypothesis underlying the industry disaggregation in the SUT and IOT. It is worth emphasising that identifying such differences in the product structure of inputs is only possible with a data-driven disaggregation of the industry-level input structures between types of firms (export-oriented and domestic market firms in our case). It is impossible to determine such technological differences when data on the product detail of purchases by firms is unavailable and the input vectors of industries in the SUT are disaggregated in proportion to industry-level totals as done in prior work (for example Chong et al. (2019); Fetzer et al. (2018); Nilsson et al. (2019)).

First results from the work on employment indicate that export-oriented firms have, on average, a lower share of total manufacturing employment than of total manufacturing value added, in other words, they have higher value added per person employed than domestic market firms. Moreover, their workforce has, on average, higher levels of educational attainment, and they pay higher wages.

([20]) Technical coefficients are the result of the normalisation of an industry's input structure by its output, in other words, they indicate how much of each product category of intermediate input is required per unit of output.

Through value chain analysis based on the extended IOT (Michel et al. (2018)), it has been shown that: (i) the difference in the estimate of vertical specialisation (import content of exports) in manufacturing between conventional and extended IOT amounts to roughly 2 percentage points; (ii) exports of export-oriented manufacturers generate substantial value added in other Belgian firms, in particular providers of services; (iii) Belgium's backward participation in global value chains is mainly due to export-oriented manufacturers and its forward participation is due to other firms, (iv) export-oriented manufacturers participate in value chains that comprise, on average, a greater number of upstream and downstream production stages and of which a greater share is located abroad.

# 4. Outstanding issues

Work on extended SUT and IOT for Belgium is ongoing. The results obtained so far have raised interest for this work. The aim is now to go beyond the exporter status criterion by investigating the data situation for the other two most commonly used criteria (ownership and size) and to produce further extended SUT and IOT for 2015. There remain a lot of issues to be addressed in this work for Belgium. In this section, we provide a structured overview of these outstanding issues. We have grouped these issues into two categories: specific definition and data issues for each criterion and cross-cutting issues.

## Definitions and data

There are specific issues regarding the definition and the underlying data for each of the three disaggregation criteria. They are summarised in Table 2. As emphasised in OECD (2015), the aim is to construct extended SUT and IOT that minimise within-industry heterogeneity respecting given confidentiality constraints, in other words, those that producers of SUT and IOT generally face when it comes to disaggregating industries, and without imposing the burden of new data collections as well as limiting the extra compilation and processing burden. Hence, the discussion here is focused on existing data sources.

The standard disaggregation according to *exporter status* divides firms in each industry into exporters and non-exporters ([21]); this can be complemented by a threshold as we have done for 2010. Thereby, we distinguish firms that are export-oriented from firms that mainly serve the domestic market (domestic market firms). The motivation for applying a threshold is to increase the homogeneity of the resulting within-industry groups. Thresholds can be defined in relative terms — exports as a share of a firm's turnover — or in absolute terms — a minimum value of total firm-level exports; a double threshold in both relative and absolute terms is also possible. In the construction of extended SUT for Denmark, Nilsson et al. (2019)

([21]) There have been other efforts to separate out specific groups of firms in SUT and IOT that are related to the exporter status criterion: in the extended tables for China, processing traders have been isolated (Koopman et al. (2012)), for Mexico, firms operating under special trade regimes have been isolated (De la Cruz et al. (2011)), and for Costa Rica, firms operating in free trade zones have been isolated (Saborio (2015)).

consider firms as export-oriented if the value of their exports is at least EUR 5 000 and exceeds 5 % of their turnover. Imposing an extra absolute threshold of a minimum of EUR 1 million of exports in addition to the 25 % relative threshold for Belgian manufacturing in 2010 would reduce the number of export-oriented manufacturers by 200 (from 2 430 to 2 230) ([22]). The underlying idea is that such thresholds avoid grouping together 'small' and 'big' exporters, which may be a source of heterogeneity ([23]).

Furthermore, in line with empirical findings in Bernard et al. (2009), one may want to specifically focus on exporters that also import as a more homogenous category. This focus on two-way traders is also suggested in Ahmad (2018) as a means of further reducing heterogeneity and it could be combined with the application of thresholds, for both exports and imports. In the 2010 data for Belgium, more than 90 % of the 2 430 export-oriented manufacturers are importers and for half of these firms the value of imports represents 25 % or more of their turnover.

Regarding the exporter status criterion, firm-level data on trade in goods and services are readily available for Belgium as they are used in the construction of conventional SUT and IOT. This favourable data situation made this criterion a natural candidate for testing the construction of extended SUT and IOT. The application of a relative threshold also requires as a complement data on turnover, which is available for Belgian firms from annual accounts, structural business statistics or value added tax (VAT) records.

So far, our disaggregation according to exporter status only considers firms that are direct exporters. Ahmad (2018; pp. 12) advocates to clearly label this fact because 'a significant share (…) of total imports and exports are made by distribution firms (wholesale and retailers)'. Nilsson et al. (2019; (pp. 16)) estimate that 'about 30 % of the Danish export of goods is exported through wholesalers'. Bernard et al. (2009) also find that trade by wholesalers accounts for a substantial part of total United States trade in goods. Identifying firms that export (a significant share of their output) through distribution firms is likely to be a difficult and work-intensive undertaking. It requires not only data on exports of wholesalers and retailers but also and most importantly data on their domestic transactions. For Belgium, the VAT transaction dataset could help. This dataset records all domestic transactions subject to VAT and is already used in the construction of conventional SUT and IOT. But even then, major methodological problems need to be addressed for a reliable identification of firms that export through wholesalers and retailers, for example how to determine whether goods delivered by domestic firms to wholesalers or retailers correspond to the goods exported by these wholesalers or retailers, or how to account for transactions not subject to VAT. When considering two-way traders, the issue also arises for imports. Moreover, it remains to be seen whether firms that export (and import) through wholesalers or retailers are technologically similar to direct exporters (two-way traders) or to non-exporters.

([22]) The 200 firms that would not be considered anymore as export-oriented accounted for less than 0.1 % of total exports.

([23]) As explained in the section on statistical work, imposing a relative threshold also mattered for industry-level sample sizes in our 2010 extended tables for Belgium given that we disaggregated the product structure of output and intermediate inputs of manufacturing industries based on data from extra questionnaires on turnover and purchases by product category that is only available for a limited number of big firms. Adding an absolute threshold may also allow to reduce the cut-off percentage of the relative threshold.

**Table 2: Overview of specific issues in terms of definition and data for the three potential disaggregation criteria (exporter status, ownership and size) for the 2015 extended SUT and IOT in Belgium**

| | Exporter status | Ownership | Size |
|---|---|---|---|
| Categories of firms | Exporters \| Non-exporters<br>Export-oriented firms \| Domestic market firms<br>Two-way traders \| Other firms | Purely domestic firms \| Domestic multinationals \| Foreign affiliates | Small and medium-sized firms (SMEs) \| Big firms<br>Independent SMEs \| Dependent SMEs \| Big firms<br>(further split of SMEs into micro, small and medium-sized firms) |
| Thresholds | Absolute (EUR million of exports)<br>Relative (% share of exports in turnover or sales)<br>Combination of absolute and relative | Participation rate (10 %-50 %) | Employment (< 250 persons)<br>Turnover (≤ EUR 50 million) or balance sheet total (≤ EUR 43 million) |
| Issues for definition and thresholds | Direct exports vs exports through wholesalers<br>Import threshold? | Domestic groups<br>Direct vs indirect ownership<br>Control vs participation rate | Independence of SMEs (foreign affiliates, participation in domestic group)<br>Thresholds from EU SME definition vs specific Belgian thresholds |
| Core data (firm level) | Trade in goods<br>Trade in services | Group structure survey (for foreign direct investment (FDI) and foreign affiliate trade statistics (FATS))<br>Annual and consolidated accounts (shareholder and affiliate structure) | Employment (social security records)<br>Turnover (annual accounts and other sources)<br>Balance sheet total (annual accounts) |
| Additional data (firm level) | Turnover (for relative threshold, annual accounts and other sources)<br>VAT transaction data (for identifying exports through wholesalers) | Commercial databases (Orbis, Amadeus, …)<br>EuroGroups Register<br>Annual accounts / social security records (for thresholds)<br>Global group structure data (for indirect links) | Data on ownership (for identifying dependent SMEs) |

The standard disaggregation according to *ownership* divides firms into three categories (see guidelines in Ahmad (2018) or Fetzer et al. (2018) for the United States, Statistics Denmark and the OECD (2017) for several Nordic countries ([24])): firms without links with firms abroad (purely domestic firms), firms with (a) foreign affiliate(s) (domestic multinationals) and firms that are part of a foreign group (foreign affiliates) ([25]). In our view, the third category should be dominant with respect to the second, in other words, a domestic firm with a foreign affiliate that is itself an affiliate of a foreign group should be part of the category of foreign affiliates.

([24]) Nilsson et al. (2019) only consider domestic and foreign-owned firms.
([25]) Further categories could be added for firms that are part of a domestic group.

Traditionally, firm A is considered as an affiliate of another firm B if B has control over A, where control means the ability to determine a firm's strategy (Eurostat (2012)). Control can be exercised directly by those holding a majority of the voting power of a firm, but effective minority control with a share of less than 50 % is also possible as is indirect control through another affiliate ([26]). Hence, control and economic ownership are not equivalent. In practice, ownership is often used as a proxy for control, in other words, a participation rate with a certain threshold; participation may be direct only or also indirect ([27]). The chosen threshold generally lies between a participation rate of 10 % (the foreign direct investment (FDI) threshold) and 50 % (majority participation).

The disaggregation with respect to ownership in the extended SUT for the United States relies on a database concerning the activities of multinational enterprises (AMNE, Fetzer et al. (2018)), while the work for Denmark is based on foreign affiliates statistics (FATS, see Nilsson et al. (2019)). Regarding data on foreign ownership for Belgium, the foremost source is the group structure survey conducted by the National Bank of Belgium (NBB), which is the basis for selecting samples for FDI and FATS surveys. This source contains information on the affiliation of Belgian firms including direct participation rates. But thresholds in terms of certain balance sheet variables are applied in the sample selection for the survey, thereby effectively excluding smaller firms even if they are domestic multinationals or foreign affiliates. For Belgium, it would be possible but likely very work-intensive to identify firms with foreign ownership links below the thresholds based on additional data. But it is not guaranteed that these below-threshold foreign affiliates and domestic multinationals are technologically similar to their bigger counterparts above the threshold. Given the workload and uncertainty about improving within-industry homogeneity, sticking to the survey thresholds may be considered the best option. Domestic multinationals and foreign affiliates identified by the survey should then be considered as exhaustive to avoid extrapolating survey results for bigger firms to smaller below-threshold firms.

For Belgium, there are indeed several other data sources that could replace or extend the sample of domestic multinationals and foreign affiliates identified from the group structure survey. The foremost of these data sources is the ownership information contained in firms' annual accounts. There is also data on shareholder structures provided in commercial databases such as Amadeus or Orbis ([28]). For the latter, the reliability of the information must be investigated and differences with respect to the group structure survey checked. Moreover, the EuroGroups Register (EGR) ([29]) is bound to improve the quality of the information on foreign affiliates.

The disaggregation of industries by *firm size may* be based on the standard definition of small and medium-sized enterprises (SMEs) in the EU as presented in the European Commission's *User guide to the SME definition* (European Commission (2015)). In this definition, a firm is considered an SME if it employs fewer than 250 persons and its turnover is less than or equal to EUR 50 million ([30]); moreover, the definition takes into account whether a firm belongs

([26]) For a full discussion, see the *Foreign AffiliaTes Statistics (FATS) Recommendation Manual*, Eurostat (2012).
([27]) Taking into account indirect participation severely raises data requirements.
([28]) For more information on these databases, see: www.amadeus.bvdinfo.com and www.orbis.bvdinfo.com.
([29]) See: https://ec.europa.eu/eurostat/web/structural-business-statistics/structural-business-statistics/eurogroups-register.
([30]) The turnover threshold may be replaced by a threshold in terms of the balance sheet total (≤ EUR 43 million), see European Commission (2015).

to a group, either domestic or foreign. A firm with below threshold values for employment and turnover is likely to be different in technological terms if it is controlled by a foreign multinational or a big domestic firm rather than being completely independent. Hence, homogeneity in terms of size also depends on whether firms belong to a domestic or international group. This has been taken into account in prior work on extended SUT and IOT with a size class disaggregation (Chong et al. (2019) for the Netherlands; Statistics Denmark and the OECD (2017) for the Nordic countries), at least for being part of an international group. The tables for these countries distinguish independent and dependent or linked SMEs.

The European thresholds for defining an SME may also be considered as too high for a small country like Belgium and lower thresholds may be defined accordingly. Moreover, the disaggregation may target more than two size classes, for example Chong et al. (2019) distinguish small and medium-sized enterprises as two separate groups.

For Belgium, data on turnover and employment is readily available and used in the construction of conventional SUT and IOT and in the national accounts. Taking group affiliation into account for the disaggregation by size class requires data on ownership. For foreign ownership, the data described above for the ownership criterion could be used. But the data from the group structure survey may prove insufficient because many SMEs that are actually foreign-owned will not be identified as such due to the underlying size-based thresholds of the survey. Hence, unless one considers this omission does not influence the technological homogeneity of the within-industry groups of firms, it becomes necessary to identify smaller foreign-owned firms based on additional sources as suggested above. Furthermore, SMEs that belong to domestic groups should then also be identified.

## Cross-cutting issues

There are many further issues faced by the statistician in the construction of extended SUT and IOT whatever the chosen criterion for industry disaggregation. How these issues are addressed will define the scope of the exercise and the associated workload.

In this context, it is important to re-emphasise that extended SUT and IOT can serve different objectives, which, in turn, matters for methodological choices. The first and foremost goal is to improve the within-industry technological homogeneity in the SUT and IOT by grouping together producers that have similar technologies (input structures) not only because they produce similar goods or services but also due to similarities in other respects such as size, ownership and exporter status. The disaggregations have so far been introduced ex-post, in other words, into already balanced and published conventional SUT. Such ex-post extended SUT and IOT give a flavour of the extent of heterogeneity and allow us to produce analytical results ([31]). However, the statistical production process would really only be altered through an ex-ante approach where such disaggregations become part of the construction process of conventional SUT and IOT. In such an ex-ante approach, the disaggregations could contribute to an improved production process and higher quality conventional tables by: (i) revealing links between firms and thereby improving the understanding of the origin of

([31]) As mentioned before, extended SUT and IOT can contribute to refining and extending value chain analysis, in particular by correcting the downward bias in the results on the import content of exports that comes from averaging import intensities over different types of producers within industries (Piacentini and Fortanier (2015)) and by addressing issues such as the distribution of income and employment within value chains (Ahmad (2018)).

discrepancies between data sources used in the construction of the SUT, and (ii) making the balancing process of the SUT smoother. Given the role of SUT as the central balancing tool of the national accounts, the ex-ante approach could also contribute to improving the quality of national accounts. Moreover, it would allow for a more reliable test of the validity of the hypothesis underlying the disaggregation, in other words, whether there are within-industry technological differences between export-oriented and other firms, between big and small firms or between domestic multinationals, foreign affiliates and purely domestic firms.

A first issue is the choice of *which industries to disaggregate*. For the 2010 extended SUT and IOT for Belgium, we restricted the exercise to manufacturing industries. This should be extended to selected service industries. But as emphasized in Ahmad (2018), it is neither feasible nor useful to disaggregate all industries. This is obviously the case for industries where all firms belong to the same group, for example if there are only non-exporters in an industry. Moreover, it is simply not meaningful to disaggregate certain industries for some criteria, for example there is no use disaggregating public administration or defence for any of the three criteria. The lack of data is another argument for not disaggregating certain industries. A list of industries selected for disaggregation should be drawn up at the start of the exercise. This list may be different for the three disaggregation criteria.

The number of industries to be disaggregated also depends on the level of industry breakdown at which the disaggregation into groups of firms is implemented. Disaggregating NACE Groups (3-digit industries) rather than NACE Divisions (2-digit industries) yields fewer firms per group within industries. Thus, a disaggregation of more aggregated industries may contribute to avoiding too small and non-representative samples for certain groups of firms within certain industries. But this implies a trade-off: working with more aggregated NACE industries reduces within-industry homogeneity, which the disaggregation exercise for the construction of extended SUT and IOT aims to increase. When working with more aggregated NACE industries, one implicitly assumes that the alternative disaggregation criteria — size, ownership, exporter status — matter more for technological homogeneity than product similarity ([32]).

At this point, it is useful to re-emphasise that the 2010 extended SUT and IOT for Belgium are based on disaggregations by exporter status at the most detailed industry level of the conventional SUT. Within-industry samples for export-oriented and domestic market firms were big enough at this level of industry breakdown for disaggregating output and intermediate input purchases. For determining product distributions, extra data for a smaller sample of firms was used and the sample size for either export-oriented or domestic market firms proved insufficient in a few industries. To address this issue, we have chosen to disaggregate the product distributions for these industries proportionally to the industry totals. Note that in prior work on extended SUT such proportionality is the rule in the calculation of product distributions for groups of firms within industries (for example Nilsson et al. (2019); Fetzer et al. (2018), Chong et al. (2019)). Proportionality yields if not identical then at least similar within-industry product distributions for output and intermediate inputs of different groups of firms, for example big firms and SMEs or exporters and non-exporters. This contradicts the originally pursued goal of revealing heterogeneity in terms of input structures between different groups of firms within industries, and it represents a problem for analyses based on extended industry-by-industry IOT.

([32]) Distinguishing different categories of firms in more aggregated industries may also contribute to avoiding confidentiality issues.

A second issue concerns the *combination of disaggregation criteria* for producing extended SUT and IOT. This is advocated in OECD (2015) and to some extent applied in Ma et al. (2015) ([33]). It would, for example, seem natural to produce extended SUT and IOT that combine the size and ownership criteria. The three criteria are likely to isolate to a large extent the same firms because firms that are part of a multinational group are mostly big and export-oriented as shown in Bernard et al. (2009) for the United States. But there are caveats from a statistical point of view. Combining disaggregation criteria leads to more groups of firms per industry and is therefore likely to give rise to problems of sample size and confidentiality. Some groups may not contain a representative number of firms but are not completely empty either. In that case, it may be good to create a group of 'other firms' that groups together those groups and is therefore relatively heterogeneous. The sample size problem is exacerbated when it comes to using surveys that cover only a restricted sample of firms like the surveys on the product detail of turnover and purchases in Belgium. Hence, the construction of extended SUT and IOT for multiple disaggregation criteria requires a careful definition of groups based on a prior analysis of samples.

The most difficult issue in deriving extended SUT is the *disaggregation of the rows* (Ahmad (2018)), in other words, determining the origin of goods and services that are purchased for domestic intermediate or final use or are exported. The origin may be imports or domestic production, and the latter may be production of the different types of firms, for example purely domestic firms, domestic multinationals or foreign affiliates. Estimating the import flow matrix (consumption of goods and services that are imported) is the easier part. As described above, we have done this for Belgium for the 2010 extended tables as follows. The conventional SUT for Belgium comprises an import flow matrix that has been estimated based on the most detailed trade data (by firm, product, transaction type, …) according to the methodology developed in Van den Cruyce (2004). It provides the relevant information for all industries that are not disaggregated as well as for domestic final use and exports (re-exports). The use of imported intermediates for different types of firms within industries, in other words, those that are being disaggregated, can be calculated with product-level import data by type of firm.

Determining the type of producing firm for the consumption of domestically-produced goods and services is a more complicated task. The extended supply table shows the estimated value of production by type of firm for each product category. This is a constraint for attributing domestic production to firm types for the different use categories. For exports of domestic origin, this attribution can be determined from data on product-level exports by firm type. For the other use categories, this attribution requires data on domestic transactions by type of firm. For the 2010 extended SUT for Belgium, we have attributed exports to export-oriented and domestic market manufacturers based on detailed export data and then distributed the remainder proportionally over firm types (for each product category). The availability of VAT transaction data for Belgium should allow for an improvement with respect to this proportionality, although this is likely to be rather work-intensive. There are nonetheless three caveats to be kept in mind for the use of this dataset.

([33]) Ma et al. (2015) construct IOT for China that not only distinguish processing exporters and other firms as in Koopman et al. (2012) but also foreign-invested and Chinese-owned firms.

(i) The VAT transaction dataset covers only transactions between firms that have to submit VAT declarations. Transactions between households and firms or between government bodies and firms are not included; hence, it could only be used for attributing most of intermediate consumption and investment. The attribution of all other domestic final use categories to firm types still needs to be done proportionally to production by firm type.
(ii) There are no product codes mentioned in the VAT transaction dataset. The type of product delivered can only be inferred from the industry code of the firm that is the supplier in the transaction. This requires matching the VAT transaction dataset with the business register which contains firms' industry codes. The product code of a transaction is nevertheless uncertain because suppliers may be wholesalers or produce more than a single product.
(iii) It remains to be seen how much difference the use of the VAT transaction dataset will actually make, given the constraints on its use for disaggregating the rows (only for intermediate consumption plus the need to respect production by firm type totals from the supply table). The question is to what extent a row disaggregation based on VAT transaction data rather than a proportionality assumption will change the results of the derivation of extended IOT and indicators based on these tables.

The last statistical issue that we discuss in detail here relates to *respecting values from the conventional SUT* as totals for disaggregated industries. This is the approach followed for the 2010 extended SUT for Belgium. It entails the application of proportional or bi-proportional RAS adjustments for what is not covered by the data underlying the disaggregation of industries and for values that have been adapted in the balancing process of the conventional SUT. As a consequence, disaggregation results may be altered with respect to what is contained in the underlying data sources. This cannot be avoided unless extended SUT are constructed from scratch with a fully-fledged balancing process. Beyond the workload implied by such an approach, discrepancies in results compared with the conventional (published) SUT would raise new issues. Moreover, such an approach goes far beyond what has been common practice up to now and what has been advocated in prior scoping contributions (OECD (2015); Ahmad (2018)).

Finally, we want to briefly mention an issue that does not need to be addressed for Belgium: the firm-establishment adjustment, which arises for countries where national accounts and SUT are constructed based on data for establishments while the disaggregation criteria concern by definition only firms. Fetzer et al. (2018) provide a detailed description of how they have dealt with this issue in the construction of extended SUT for the United States. This is not an issue for extended SUT for Belgium since the Belgian national accounts and SUT are based on legal units, which we have referred to as firms, rather than establishments.

# 5. Conclusions

Extended SUT and IOT are an important statistical building block for improving the measurement of economic activities in times where strong global interactions make this measurement increasingly complicated. Work on extended SUT and IOT for Belgium for the year 2010 as a test case has shown that the construction of such tables is feasible for Belgium. Analysis based on these tables has produced valuable insights on technological differences and value chain integration of export-oriented and domestic market manufacturing firms. These achievements are an incentive to pursue efforts: to construct extended SUT and IOT for Belgium for other years, in particular the now most recent I-O reference year 2015, and to consider not only exporter status but also the other most common disaggregation criteria, in other words, size and ownership. This paper has provided a discussion of statistical issues to be addressed in the construction of these extended SUT and IOT.

There are specific issues in terms of definitions and data for the three disaggregation criteria that have been covered (exporter status, ownership and size) and there are more general issues that pertain to all three of them. The main cross-cutting issues are: (i) the choice of industries to be disaggregated and the industry-level breakdown at which to perform the disaggregation, (ii) whether to combine disaggregation criteria, (iii) the disaggregation of the rows in the SUT, in other words, the identification of the origin of consumed goods and services, and (iv) the consequences of adjusting results to respect the conventional SUT. Moreover, the discussion reveals that a proportional disaggregation of the product distributions for industries' output and intermediate consumption (columns of the SUT) is problematic for deriving extended IOT from the extended SUT because it leads to identical technical coefficients for within-industry groups of firms and therefore casts doubts on the relevance of the derived value chain indicators.

A few issues have not been covered because they have already been discussed *in extenso* in prior work (Michel et al. (2018)): (i) differences in the country distribution of trade between within-industry groups of firms, and (ii) the integration of extended SUT or IOT into global multi-country IOT. Further issues are bound to arise through additional analytical goals. The analysis of the distribution of income in value chains requires a disaggregation of value added components, which is likely to be particularly challenging for the operating surplus and consumption of fixed capital. The relevance of the comparison of extended SUT and IOT and derived indicators for different years depends on the stability of methods and samples over time. Finally, productivity analysis would require separate price data for within-industry groups of firms.

Extended SUT and IOT are a field with a high potential for demonstrating the policy relevance of statistical work. But a lot of issues remain to be addressed in the construction of these tables. Our aim is to pursue the work on Belgian extended SUT and IOT and explore the possibilities of the individual firm-level databases available to us in order to contribute to statistical developments in this field and to economic analyses based on these tables.

# References

Ahmad, N. (2018), 'Accounting for Globalisation: Frameworks for Integrated International Economic Accounts', forthcoming in *The Challenges of Globalization in the Measurement of National Accounts*, Ahmad, N., B. Moulton, J. D. Richardson and P. van de Ven, eds., National Bureau of Economic Research.

Bernard, A. B., J. B. Jensen and P. K. Schott (2009), 'Importers, Exporters, and Multinationals: A Portrait of Firms in the U.S. that Trade Goods', in *Producer Dynamics: New Evidence from Micro Data*, Dunne, T., J. B. Jensen and M. J. Roberts, eds., Chapter 14, pp. 513-552, University of Chicago Press.

Chong, S., R. Hoekstra, O. Lemmers, I. Van Beveren, M. Van Den Berg, R. Van Der Wal and P. Verbiest (2019), 'The role of small- and medium-sized enterprises in the Dutch economy: an analysis using an extended supply and use table', *Journal of Economic Structures*, Volume 8, Issue 8, pp. 1-24.

De la Cruz, J., R. Koopman, Z. Wang and S. J. Wei (2011), 'Estimating Foreign Value-Added in Mexico's Manufacturing exports', Office of Economics Working Paper No. 2011-04A, U.S. International Trade Commission.

Eaton, J., S. Kortum, and F. Kramarz (2004), 'Dissecting Trade: Firms, Industries, and Export Destinations', *The American Economic Review*, Volume 94, No. 2, pp. 150-154.

European Commission (2015), '*User guide to the SME definition*', Luxembourg.

Eurostat (2008), '*Eurostat Manual of Supply, Use and Input-Output Tables*', Luxembourg.

Eurostat (2012), '*Foreign AffiliaTes Statistics (FATS) Recommendation Manual*', Luxembourg.

Fetzer, J., T. Highfill, K. Hossiso, T. F. Howells III, E.H. Strassner and J. A. Young (2018), 'Accounting for Firm Heterogeneity within U.S. Industries: Extended Supply-Use Tables and Trade in Value Added using Enterprise and Establishment Level Data', forthcoming in *The Challenges of Globalization in the Measurement of National Accounts*, Ahmad, N., B. Moulton, J. D. Richardson and P. van de Ven, eds., National Bureau of Economic Research.

FPB (2015), '*Tableaux Entrées-Sorties 2010 - SEC 2010 / Input-outputtabellen 2010 - ESR 2010*', Federal Planning Bureau, Brussels.

Koopman, R., Z. Wang and S.-J. Wei (2012), 'Estimating domestic content in exports when processing trade is pervasive', *Journal of Development Economics*, Volume 99, Issue 1, pp. 178-189.

Ma, H., Z. Wang and K. Zhu (2015), 'Domestic content in China's exports and its distribution by firm ownership', *Journal of Comparative Economics*, Volume 43, Issue 1, pp. 3-18.

Michel, B., C. Hambÿe, and B. Hertveldt (2018), 'The Role of Exporters and Domestic Producers in GVCs: Evidence for Belgium based on Extended National Supply and use Tables Integrated into a Global Multiregional Input-Output Table', National Bureau of Economic Research Working Paper No. 25155, forthcoming in *The Challenges of Globalization in the Measurement of National Accounts*, Ahmad, N., B. Moulton, J. D. Richardson and P. van de Ven, eds., National Bureau of Economic Research.

Miroudot, S. (2016), 'Global Value Chains and Trade in Value-Added: An Initial Assessment of the Impact on Jobs and Productivity', *OECD Trade Policy Papers*, No.°190, OECD Publishing, Paris.

Moulton, B. and P. van de Ven (2018), 'Addressing the Challenges of Globalization in National Accounts', forthcoming in *The Challenges of Globalization in the Measurement of National Accounts*, Ahmad, N., B. Moulton, J. D. Richardson and P. van de Ven, eds., National Bureau of Economic Research.

Nilsson, M., J. van der Kamp, N. Mortensen, and P. Jensen (2019), 'Extended Supply and Use Tables with Applications', report prepared for Eurostat, Statistics Denmark.

OECD (2015), Terms of Reference, Expert Group on Extended Supply-Use Tables, Paris.

Piacentini, M. and F. Fortanier (2015), 'Firm heterogeneity and trade in value added', prepared for OECD Working Party on International Trade in Goods and Trade in Services Statistics, STD/CSSP/WPTGS(2015)23, OECD Publishing, Paris.

Saborio, G. (2015), 'Costa Rica: An Extended Supply-Use Table', paper prepared for 23rd IIOA Conference, Mexico City.

Statistics Denmark and the OECD (2017), 'Nordic Countries in Global Value Chains', Copenhagen.

Timmer, M. P., E. Dietzenbacher, B. Los, R. Stehrer and G. J. de Vries (2015), 'An Illustrated User Guide to the World Input–Output Database: the Case of Global Automotive Production', *Review of International Economics*, Volume 23, Issue 3, pp. 575-605.

United Nations (2018), 'Handbook on Supply, Use and Input-Output Tables with Extensions and Applications', Department of Economic and Social Affairs, United Nations, New York.

Van den Cruyce, B. (2004), 'Use Tables for Imported Goods and Valuation Matrices for Trade Margins — an Integrated Approach for the Compilation of the Belgian 1995 Input-Output Tables', *Economic Systems Research*, Volume 16, Issue 1, pp. 33-61.

# Data appendix

**Figure A.1:** **Industry-level value-added to output ratios for export-oriented and domestic market firms, manufacturing, Belgium, 2010**
(%)

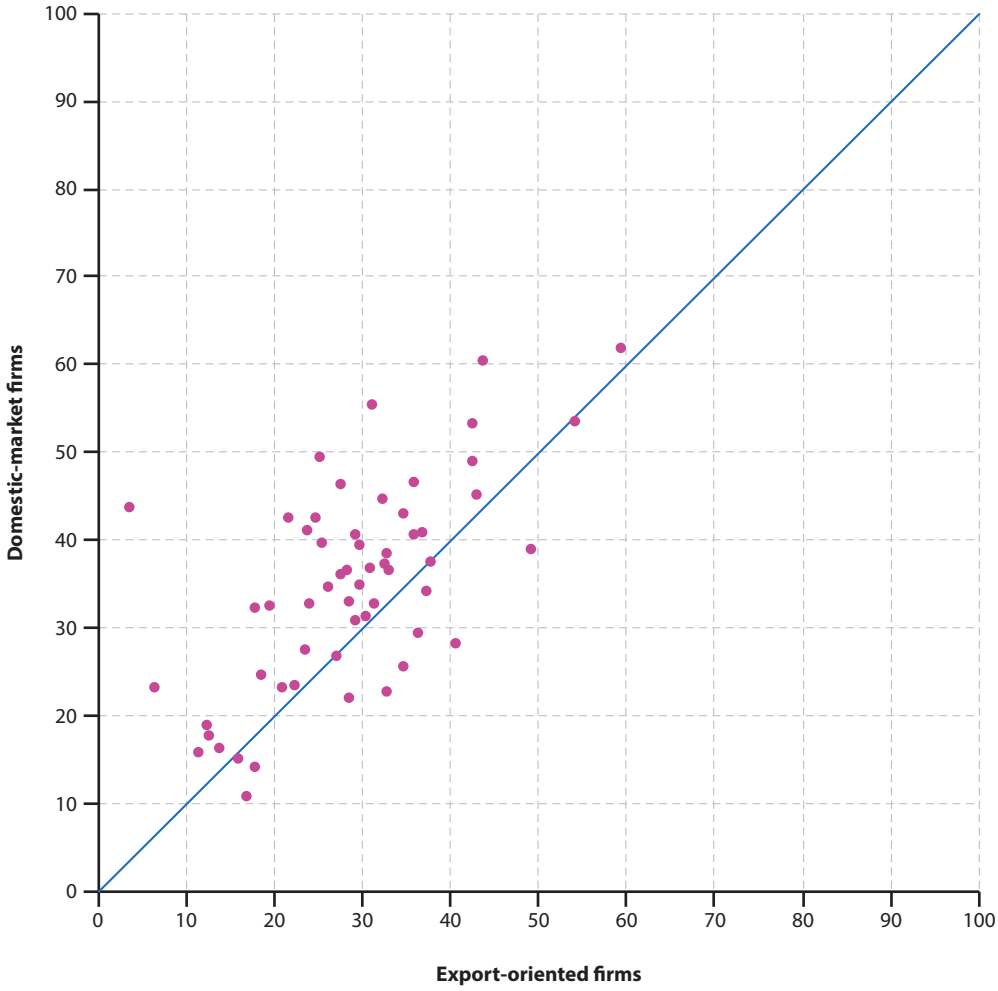
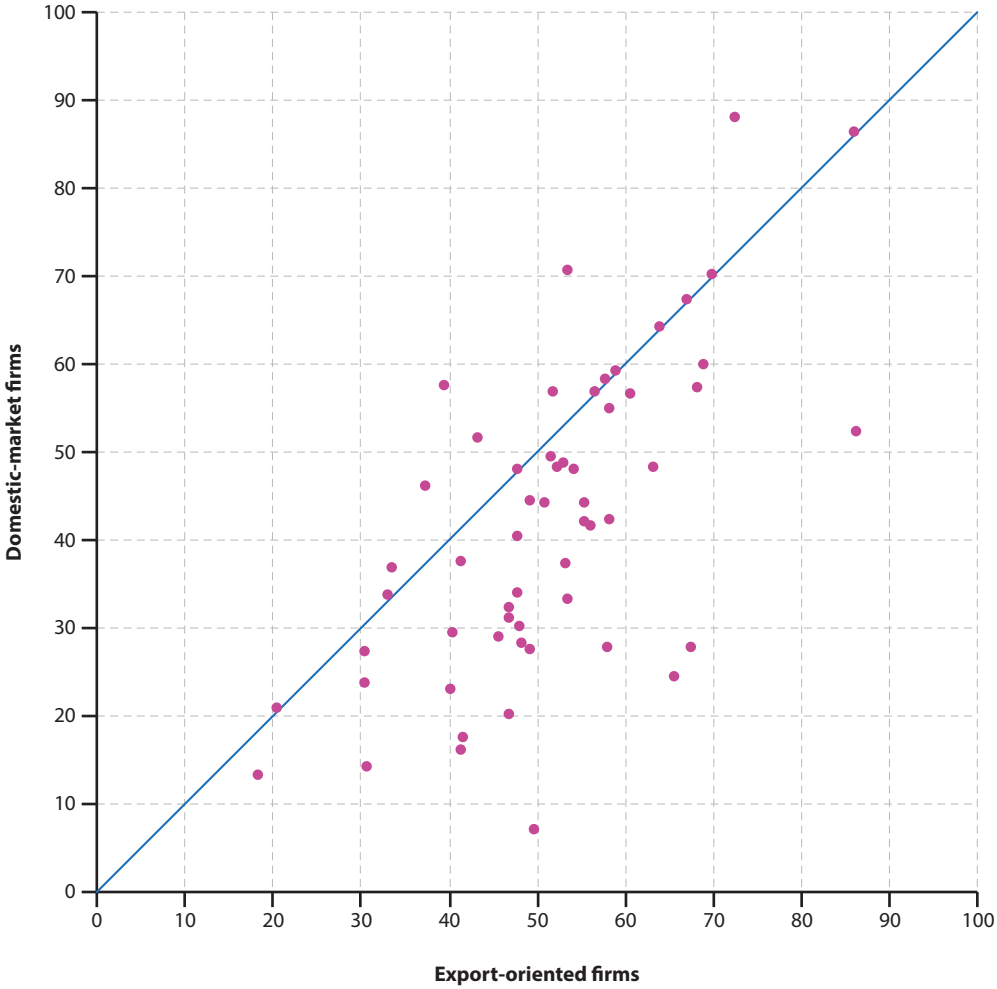


**Figure A.2: Industry-level shares of imports in total inputs for export-oriented and domestic market firms, manufacturing, Belgium, 2010**
(%)



# 3 An alternative hedonic residential property price index for Indonesia using big data: the case of Jakarta ([1])

ARIEF NOOR RACHMAN ([2])

**Abstract**: Monitoring property price dynamics is a necessary task for central banks in order to maintain financial stability in the economy. Big data offers potential as a new source of data that might be used to produce official statistics on property. In this paper, we develop an alternative residential property price index (RPPI) for the secondary market for houses from online residential property listings using the time-dummy hedonic regression method. The dataset is based on residential property advertisement listings from Indonesia's major property web portals from January 2016 to September 2018. For this prototype index, the study initially focuses on Jakarta, the capital city of Indonesia. Our regression outputs generally show promising results and have the potential to become an official housing index. Future development will extend the index coverage to other large cities in the country and improve the characteristic variables in the model.

**JEL codes**: C43, E30, R31

**Keywords**: residential property prices index, big data, hedonic regression time-dummy method

([1]) Revised version of a paper that was presented to Eurostat's International Conference on Real Estate Statistics in Luxembourg, 20-22 February, 2019.

([2]) Statistics Department of Bank Indonesia.

# 1. Introduction

Most central banks monitor property prices as part of their work to maintain national financial stability. Therefore, Bank Indonesia (BI) has established a residential property prices survey and commercial property prices survey on a quarterly basis. The bank has been conducting two kinds of surveys for residential property statistics, both for the primary market (newly-built houses) and the secondary market (used houses or second-hand houses) with different methods to compute a residential property price index (RPPI). The primary market RPPI is computed using a chained index method based on the list price of new houses as provided by key property developers in 16 major cities, whereas appraisal methods are used to compute an RPPI for the secondary market in 10 major cities.

The compilation of an RPPI is a tricky process. A number of problems may arise across different stages of the process, from the identification of data sources to index calculation methods (Eurostat (2013)). The identification of a reliable data source is an issue that arises when computing an RPPI for Indonesia. Data on declared property transactions such as administrative data from the land registry or property tax records are difficult to acquire. The decentralisation of the administration for property taxes to local or municipal government makes it more difficult to collect data on transactions due to non-standard data records, wider coverage and reluctance on the part of local government to share data. These kinds of problem have led Bank Indonesia to conduct property price surveys using list prices from developers and appraisals of the activity of real estate agents as alternatives to the compilation of an RPPI. Timeliness is one of the main benefits when using asking (or listed) prices to construct property price indices. Nevertheless, this approach can potentially be a major weakness, insofar as differences between asking prices and actual transaction prices may result in misleading estimates. However, an RPPI based on asking prices may be considered a feasible solution for monitoring purposes, especially in the absence of data for actual transactions (IMF (2018)). Lyons (2019) also found that asking/listed prices can be an accurate indicator of actual transaction prices in Ireland's market for houses.

Recent developments of digital data known as 'big data' offer opportunities and benefits to official statistical institutions, such as: the ability to produce new indicators; the possibility to bridge time lags for existing official statistics; and the provision of an alternative source of data to produce official statistics. According to Hammer et al. (2017), big data is defined as a by-product of business and administrative systems, social networks and the internet of things and is often characterised by its high-volume, high-velocity and high-variety (3Vs) of data. However, big data also provides several challenges that need to be overcome, such as: i) concerns over data quality ([3]); ii) ensuring (legal) access to the data, considering big data is typically owned by private entities; and iii) developing advanced skills and making available the necessary technology to make use of such data (Das et al. (2014) and Hammer et al. (2017)). Therefore, statistical institutions need to be careful when using big data as a new source of official statistics.

This paper develops an alternative RPPI, making use of big data by drawing on online advertisements for property. The use of big data has the potential to improve the compilation of the RPPI and to challenge the existing practices employed to compute the existing RPPI.

([3]) This concerns having to deal with large volumes of data that need to be processed, cleaned/verified and then summarised without losing too much information.

Online data based on the listed price of properties as found in advertisements offers an immediate, inexpensive, and considerable amount of (alternative) data for constructing an RPPI. This study employs a direct hedonic approach to calculate robust property price indices based on the availability of data for various property characteristics. As a prototype index, the coverage for this study is limited to the Indonesian capital city, Jakarta.

This paper is organised as follows. The first section provides an introduction detailing the background to this study. Section two explains the data and the methodology used. A discussion of the results is presented in section three. Finally, conclusions and further work are presented in section four.

# 2. Data and methodology

## 2.1. Data sources

In the development of an alternative RPPI using big data, we collected monthly data from the two largest web portals for property advertisements in Indonesia, which together account for more than 50 % of the total market. Bank Indonesia secured the acquisition of these data through non-disclosure agreements (NDAs) with the two websites. The preparation and extraction of data from the two web portals was organised using virtual machines and Hadoop software (more details concerning the steps taken are presented at the end of this article in a *Data appendix*).

The data used were individual listings/advertisements for property with the following attributes: initial asking (or listed) price, offer type (for sale or rent), property type, lot size, dwelling size, number of bedrooms, number of bathrooms, address, and additional characteristics which are recorded as a 'free-text' description, such as the presence (or not) of a garage, gated property, swimming pool, its specific location or its distance from public facilities, and so on. For simplicity purposes, these 'free-text' characteristics were left aside in this study because the information was often incomplete and too granular to extract.

As it was initially an experiment, the study only focused on listings of houses that are 'for sale', while other types of residential property such as apartments/flats were excluded and left for future developments. Furthermore, the study only included the first instance of any advertisement/listing for each property; as such, only listings from the first month that an advertisement appeared were included, unless the listed price subsequently changed. Taking the listed price of each property on a monthly basis would implicitly give a larger weight to those properties which take longer to sell.

As mentioned above, the study was limited to computing an RPPI only for Jakarta, the capital city of Indonesia. Jakarta is one of the biggest cities in the world with a population of around 10 million people. As the nation's capital and the city with the largest population in Indonesia, Jakarta has the highest number of property transactions in Indonesia (when compared with other cities). Jakarta is believed to account for around one third of the national property market. Our dataset shows that Jakarta accounted for around 36 % of all online listings at a national level.

The study also used another dataset as a set of weights when aggregating the hedonic indices for each district into a composite property price index for houses across the whole of Jakarta. The total value of mortgage collateral by regions/district was used as a proxy for all transactions, in the absence of data covering both cash and mortgage-financed transactions. This dataset consists of individually appraised collateral values for mortgage loans that are derived from the centralised banking debtor information system which is jointly managed by Bank Indonesia and the Indonesian Financial Services Authority (OJK). We found that the share of each district in the total value of mortgage collateral value was relatively closely related to the share of each district in the total number of observations; in both cases South Jakarta had the biggest share among the five districts that compose the Indonesian capital. A table with this comparison is presented in the *Data appendix*.

## 2.2 Data treatment

Cleaning the data set was a crucial step when dealing with sources of big data due to concerns over data quality (as mentioned above). Since we obtained data from individual listings on web portals there were several issues regarding the data, including: (1) human error in data entry; (2) non-standard addresses — as a free-text field is employed; and (3) duplicate advertisements (which are mainly caused by the fact that one property can be advertised by more than one seller in a single portal as well as across different portals, and advertisements tend to be re-posted after their initial expiration date if the property has not been sold). When preparing the dataset, we removed all duplicate records and any data considered to be corrupt/incomplete.

The next step was to make statistical edits based on the assumption of a normally distributed dataset. We removed spurious price values using a median absolute deviation (MAD) test on price per unit of property size; the same method was also applied for building size. We removed those observations where the lot size was greater than 600 $m^2$, and also deleted any observations where the number of bedrooms was greater than 10 and the number of bathrooms was greater than eight, based on histogram (or bell-shaped curve) patterns. We trimmed any data cells which lay outside of the observed bell curve tails (outliers).

Finally, we ran a preliminary regression to identify further outliers using the Cook's distance method. This method identified outliers based on the combination of each observation's leverage and residual values, with the following formula:

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{(p+1)\hat{\sigma}^2}$$

Where $D_i$ is Cook's distance for observation *i*, $\hat{Y}_j$ is the fitted response value, $\hat{Y}_{j(i)}$ is the fitted response value obtained when observation *i* is removed, $\hat{\sigma}^2$ is the mean squared error of the regression, and *p* is number of predictors. We removed those observations with value of D that was greater than $4/n$, a conventional standard (see O'Hanlon (2011)). Table 1 presents a record of the preparation steps applied for each district of Jakarta. On average we identified (and removed) around 5 % of records as outliers using Cook's distance method, while an additional 25 % of the data was removed due to statistical edits.

**Table 1:** Data records

| Year | Steps | Districts | | | | |
|---|---|---|---|---|---|---|
| | | West Jakarta | Central Jakarta | South Jakarta | East Jakarta | North Jakarta |
| 2016 | # obs of active listings | 618 858 | 110 293 | 852 252 | 457 734 | 369 972 |
| | # obs of first instance listings (1) | 77 031 | 14 527 | 103 925 | 55 930 | 42 378 |
| | # obs raw data (2) | 40 141 | 7 357 | 59 922 | 30 625 | 20 370 |
| | # obs after stat edits | 32 868 | 4 946 | 41 541 | 23 421 | 15 474 |
| | # obs (3) | 31 184 | 4 662 | 39 626 | 22 180 | 14 640 |
| | # obs percentage from raw rata (%) | **77.7** | **63.4** | **66.1** | **72.4** | **71.9** |
| | # obs percentage from stat edits (%) | **94.9** | **94.3** | **95.4** | **94.7** | **94.6** |
| 2017 | # obs of active listings | 579 507 | 106 175 | 905 145 | 450 882 | 319 099 |
| | # obs of first instance listings (1) | 82 150 | 14 159 | 130 380 | 67 901 | 39 626 |
| | # obs raw data (2) | 38 477 | 7 650 | 66 294 | 36 237 | 18 498 |
| | # obs after stat edits | 31 264 | 5 046 | 43 575 | 28 921 | 14 795 |
| | # obs (3) | 29 831 | 4 762 | 41 647 | 27 426 | 14 026 |
| | # obs percentage from raw rata (%) | **77.5** | **62.2** | **62.8** | **75.7** | **75.8** |
| | # obs percentage from stat edits (%) | **95.4** | **94.4** | **95.6** | **94.8** | **94.8** |
| 2018 (9 months) | # obs of active listings | 403 232 | 78 728 | 657 112 | 348 836 | 225 357 |
| | # obs of first instance listings (1) | 64 024 | 14 666 | 151 420 | 88 366 | 45 423 |
| | # obs raw data (2) | 23 743 | 4 673 | 37 877 | 30 272 | 14 163 |
| | # obs after stat edits | 20 114 | 3 516 | 29 619 | 20 645 | 10 998 |
| | # obs (3) | 19 153 | 3 331 | 28 231 | 19 509 | 10 370 |
| | # obs percentage from raw rata (%) | **80.7** | **71.3** | **74.5** | **64.4** | **73.2** |
| | # obs percentage from stat edits (%) | **95.2** | **94.7** | **95.3** | **94.5** | **94.3** |

(1) Only new advertisements (ads) in each month, removing repeated advertisements.
(2) Raw data after removing the duplicated advertisements and corrupted data.
(3) Clean data after statistical edits and outliers are removed using the Cook's distance method.

## 2.3 Methodology

As mentioned in the Introduction, the calculation of an RPPI is a complex process because houses are infrequently sold and heterogeneous in terms of their structural characteristics such as location, size and facilities. This may lead to quality issues for price measurements since the differences in housing characteristics are hard to control, especially with a limited frequency of transactions. To identify factors for quality changes, quality-mix adjustments are needed to avoid misleading interpretations of the resulting indices. Silver (2016) identified several methods for making quality-mix adjustments such as hedonic methods, repeat sales, sales price appraisal ratios, and so on. The hedonic method is believed to be more preferable when compared with repeat sales due to its ability to use data on relevant property characteristics using regression techniques (Hülagü et. al (2015)). Furthermore, hedonic

regression analysis of house prices decomposes the overall price and provides estimates of marginal value for each of the characteristics. Li et al. (2006) highlighted three main approaches for hedonic methods: the time-dummy approach; the characteristics approach; and the hedonic (price) imputation approach. For a better explanation of these methods, see Diewert (2003), Hill (2012) and Silver (2016).

Silver (2016) indicated that both the characteristics and hedonic imputation approaches have major advantages over the time-dummy approach, but for simplicity we decided to continue with the time-dummy approach since it can immediately derive a price index from the estimated time-dummy coefficients. Li et al. (2006) also mention that the imputation method would likely give the same result as the hedonic method given the same dataset.

Our model specification used the semi-log regression model since the variable for house prices (in levels) was not normally distributed (there was a positively skewed distribution). The basic semi-log hedonic model is represented as follows:

$$\ln p_n^t = \beta_0^t + \sum_{\tau=1}^{T} \delta^\tau D_n^\tau + \sum_{k=1}^{K} \beta_k^t z_{nk}^t + \varepsilon_n^t$$

Where:

$p_n^t$ is the price of property *n* at time *t*;

$z_{nk}^t$ is k characteristic variables of property *n* at time *t*;

$\beta_0$ and $\beta_k$ are intercepts and house characteristic parameters; and

$\delta^\tau$ is a dummy coefficient.

Our hedonic model only has quantitative characteristics such as building size, lot size, number of bedrooms and number of bathrooms. The number of bedrooms and the number of bathrooms are treated as dummy variables. We have three dummy variables for the number of bedrooms — one and two bedrooms, three bedrooms, and greater than four bedrooms (four bedrooms is used as a reference based on the highest frequency for the number of bedrooms). We also have three dummy variables for the number of bathrooms (with three bathrooms as a reference).

One of the shortcomings of this model specification is the lack of other characteristic variables that may be important (such as locational advantage, the condition of the property/house, or the age of the building). However, we had difficulties to identify the location advantage of property because the address and important information about the location were often incomplete, had less detail or were composed of 'free-text' information. The main work done so far in this area was centred on identifying property locations at the district level. On the other hand, information about building age or major renovation records were generally not disclosed in the advertisements. In order to identify the location advantage, we stratified the calculation of indices into five districts that together formed the metropolitan area of Jakarta, in other words, Central Jakarta, North Jakarta, East Jakarta, South Jakarta and West Jakarta.

To calculate an RPPI, we followed the methodology/calculations employed for the Japanese residential property price index (JRPPI) (Land Economy and Construction Industries Bureau (2016)), using the rolling window technique to compute an RPPI from the hedonic regression time-dummy method. Estimated time-dummy coefficients ($\hat{\delta}^{\tau}$) were arranged as follows:

**Table 2:** Rolling window technique for the compilation of time-dummy coefficients

| *Regression r* | *1* | *2* | *3* | *…* | | | *…* | $T-\tau+1$ | *…* | $T$ |
|---|---|---|---|---|---|---|---|---|---|---|
| *1* | $\hat{\delta}_1^1$ | $\hat{\delta}_1^2$ | $\hat{\delta}_1^3$ | … | $\hat{\delta}_1^{\tau}$ | | | | | |
| *2* | | $\hat{\delta}_2^2$ | $\hat{\delta}_2^3$ | … | $\hat{\delta}_2^{\tau}$ | $\hat{\delta}_2^{\tau+1}$ | | | | |
| *3* | | | $\hat{\delta}_3^3$ | … | $\hat{\delta}_3^{\tau}$ | $\hat{\delta}_3^{\tau+1}$ | … | | | |
| *…* | | | | … | | … | … | | … | |
| $T-\tau+1$ | | | | | | | | $\hat{\delta}_{T-\tau+1}^{T-\tau+1}$ | … | $\hat{\delta}_{T-\tau+1}^{T}$ |

The index can be obtained by:

$$\frac{p^{\tau+1}}{p^1}=\exp\left(\hat{\delta}_1^{\tau}\right)\times\frac{\exp\left(\hat{\delta}_2^{\tau+1}\right)}{\exp\left(\hat{\delta}_2^{\tau}\right)}$$

Suppose the base period is the first period, then the price difference between the price at period 1 and the price at period $\tau+1$ can be obtained based on the time-dummy parameter calculated for the last period (time $\tau$) of the first window time range and the time dummy parameters for the last period and the second to last periods of the next window time range (time $\tau$ and time $\tau+1$). By sequentially conducting the aforementioned calculations for all window time ranges, quality-adjusted price indices may be obtained for all time windows.

As done for the compilation of the JRPPI, the length of the window time was set to one year (12 months) which is common for analysis, as illustrated by Silver (2016). We assumed that this window would allow us to capture the seasonal dynamics of the market for houses. To compile a composite RPPI for the whole of Jakarta, we aggregated the indices for all five districts using total collateral mortgage values as weights; these data were provided by banks in 2017. Collateral mortgage values were used as a proxy for property transaction values in the absence of a more representative measure for the structure of the property market, such as tax revenues from property transactions.

# 3. Results

## 3.1. Regression results

The information used in this study indicated that house prices (in levels) were not normally distributed. Hence, we transformed the variable for house prices into a logarithmic format before running semi-log hedonic regression models. We ran a 12-month rolling windows regression from January 2016 to September 2018 for five different districts in Jakarta with a total of 110 regressions.

We ran two stages of regression, the first was to identify outliers using Cook's distance and the second regression (without outliers) to produce the index. We present a sample set of results below for a 12-month rolling window regression from January 2016 to December 2016 in North Jakarta and South Jakarta (see Table 3). These results show a relatively high degree of explanatory power as indicated by the adjusted R-square values. Given the limited availability of variables for house characteristics, such a high degree of explanatory power probably implies a relatively homogenous market for houses in these districts.

**Table 3: Hedonic regression results for North Jakarta and South Jakarta**

| | North Jakarta | | | South Jakarta | | |
|---|---|---|---|---|---|---|
| **Dependent variable: Ln Price** | | | | | | |
| **Independent variables** | **Estimates** | **Robust standard error** | | **Estimates** | **Robust standard error** | |
| Intercept | 21.2900 | 0.00851 | *** | 20.8540 | 0.00980 | *** |
| Building size | 0.0010 | 0.00002 | *** | 0.0023 | 0.00002 | *** |
| Lot size | 0.0045 | 0.00003 | *** | 0.0031 | 0.00002 | *** |
| Dum_# of bedroom 1-2 | -0.2153 | 0.00824 | *** | -0.1042 | 0.00507 | *** |
| Dum_# of bedroom 3 | -0.0440 | 0.00456 | *** | -0.4461 | 0.00983 | *** |
| Dum_# of bedroom >4 | -0.0400 | 0.00536 | *** | -0.0929 | 0.00545 | *** |
| Dum_# of bathroom 1 | -0.2225 | 0.00879 | *** | -0.2965 | 0.01033 | *** |
| Dum_# of bathroom 2 | -0.1732 | 0.00478 | *** | -0.1389 | 0.00564 | *** |
| Dum_# of bathroom >3 | 0.0082 | 0.00492 | * | -0.0095 | 0.00502 | * |
| Dum_period 2016:2 | 0.0007 | 0.00867 | | -0.0132 | 0.01039 | |
| Dum_period 2016:3 | 0.0006 | 0.00888 | | 0.0164 | 0.01023 | |
| Dum_period 2016:4 | 0.0034 | 0.00913 | | -0.0175 | 0.01053 | * |
| Dum_period 2016:5 | -0.0203 | 0.00765 | *** | 0.0003 | 0.00921 | |
| Dum_period 2016:6 | -0.0123 | 0.00971 | | 0.0390 | 0.01094 | *** |
| Dum_period 2016:7 | -0.0132 | 0.00936 | | 0.0015 | 0.01085 | |
| Dum_period 2016:8 | 0.0022 | 0.00954 | | 0.0148 | 0.01070 | |
| Dum_period 2016:9 | -0.0143 | 0.01008 | | 0.0307 | 0.01086 | *** |
| Dum_period 2016:10 | -0.0169 | 0.00829 | ** | 0.0268 | 0.00974 | *** |
| Dum_period 2016:11 | -0.0307 | 0.00916 | *** | -0.0158 | 0.00991 | |
| Dum_period 2016:12 | -0.0435 | 0.01056 | *** | -0.0030 | 0.01052 | |
| Adjusted R-squared | 0.863 | | | 0.783 | | |
| F-statistics | 4 866 | | | 7 507 | | |
| Number of observations | 14 640 | | | 39 626 | | |

Note: *** significance at 1 %, ** significance at 5 % and * significance at 10 %.

Across all five districts of Jakarta, all characteristic variables in the model were statistically significant, stable, and in line with *a priori* expectations over time ([4]). On average, building size and lot size had a positive impact on house prices, while the number of bedrooms and the number of bathrooms had mixed results. Given no change in any other characteristics, it seemed that as the number of bedrooms increased beyond four this had a negative impact on house prices as it could reduce the living space available in the remainder of the house. The same argument applied to the number of bathrooms in South Jakarta, while for North Jakarta this result was in line with the *a priori* expectations. The results also implied that a house with an additional 10 $m^2$ of lot size would be 4.5 % more expensive than average (if other variables were kept constant).

Regression results for the South Jakarta district (based on a larger number of observations) provided similar findings. The explanatory power dropped slightly but still remained relatively high, with an adjusted R-squared value of 0.78. All of the house characteristics were significant, building size had twice the impact (compared with the results for North Jakarta) while the impact of lot size in South Jakarta was slightly less significant.

The Breusch-Pagan test was applied to detect heteroscedasticity. The result showed that heteroscedasticity was present in the model. Since heteroscedasticity only affects standard errors and the coefficients remained unbiased, we calculated robust standard errors to improve the value of the t-statistic.

## 3.2 Indices

Adopting the rolling window method, we estimated time-dummy regression coefficients for an RPPI for each of the five districts that compose the Indonesian capital city. The monthly indices suffered from short-term volatility, thus we employed three-month moving averages to smooth out the series ([5]). Thereafter, we compared the new indices with the existing RPPI (based on the appraisal method); note that the existing indices were expanded from a quarterly to a monthly frequency by simply putting the same index value for each month within a specific quarter.

([4]) We compared these regression results with the regression window for one year ahead (January 2017-December 2017) and had relatively consistent signs and coefficients for each explanatory variable.
([5]) The Central Statistics Office (CSO) of Ireland publishes a national house price index using the same technique to smooth out short-term volatility; see O'Hanlon (2011).

Figure 1 shows the index for North Jakarta. In this example, the hedonic index moves in a different direction to the existing RPPI for the sample horizon. The hedonic index generally revealed that prices were falling from 2016 to early 2017 before stabilising and remaining close to their new level. By contrast, the existing RPPI showed a generally upward trend for property prices during the sample horizon, with prices accelerating at a faster pace in the second half of 2017 and early 2018.

The situation in South Jakarta was different insofar as the hedonic index initially followed a similar pattern of development to that displayed for the existing RPPI; thereafter, the hedonic RPPI increased at a more rapid pace from the last quarter of 2017 to the third quarter of 2018 and therefore stood at a higher level than the existing RPPI (see Figure 2).

The hedonic index for Central Jakarta showed a different pattern (see Figure 3). The series displayed a high degree of volatility and even when smoothed (three-month moving average) the high degree of volatility persisted. This was probably affected by the small number of observations that were available for estimation (as Central Jakarta accounted for only 4 % of the total number of observations in the whole of Jakarta).

Figure 4 (overleaf) presents a composite index for the whole of Jakarta (an aggregate covering all five districts together). This hedonic index provided promising results and less volatile results. It showed a smooth and increasing trend during the sample horizon and one which was in line with both of the existing indices — for the primary and secondary house markets — along the sample horizon.

Annual growth rates for the hedonic index were also seen to follow a similar pattern of development to that displayed for the RPPI for the secondary market for houses (based on appraisals of activity among estate agents (see Figure 5 overleaf)).

**Figure 1: Comparison of indices for the secondary market for houses, North Jakarta, January 2016-September 2018**
(January 2016 = 100)

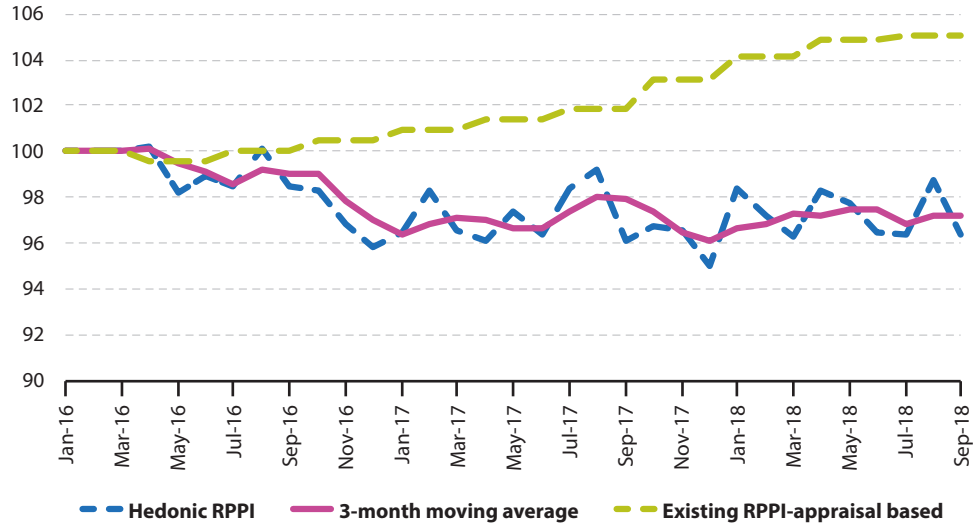


**Figure 2:** Comparison of indices for the secondary market for houses, South Jakarta, January 2016-September 2018
(January 2016 = 100)

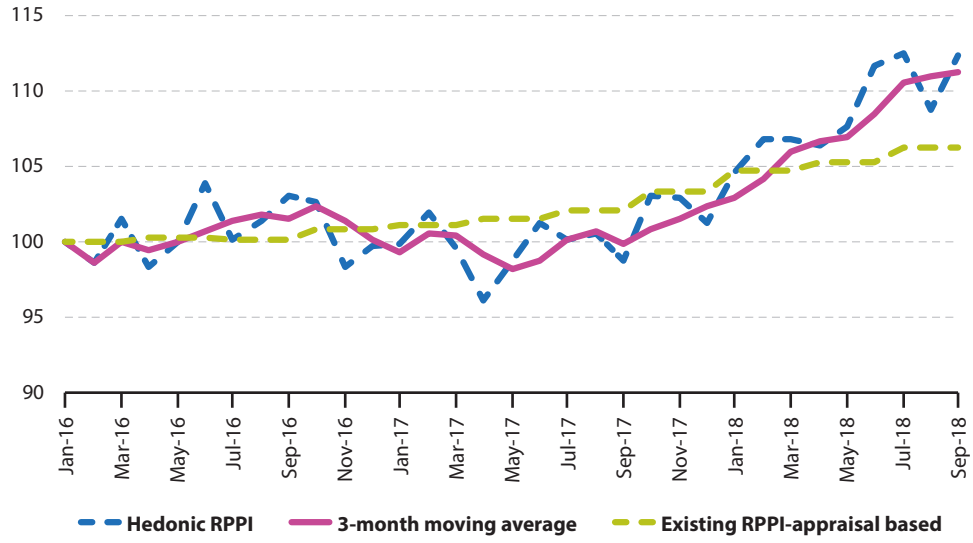



**Figure 3:** Comparison of indices for the secondary market for houses, Central Jakarta, January 2016-September 2018
(January 2016 = 100)

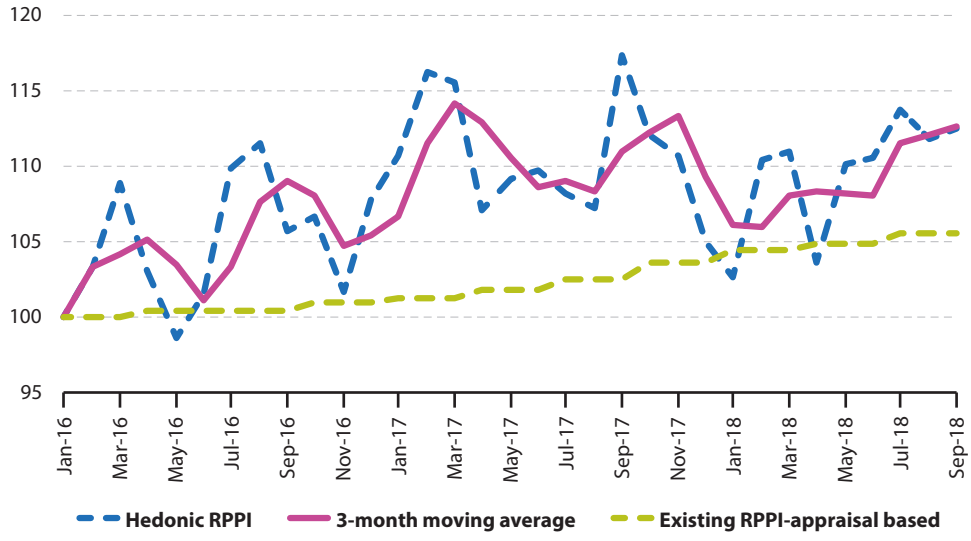


**Figure 4:** Comparison of indices for total housing markets, all of Jakarta, January 2016-September 2018
(January 2016 = 100)

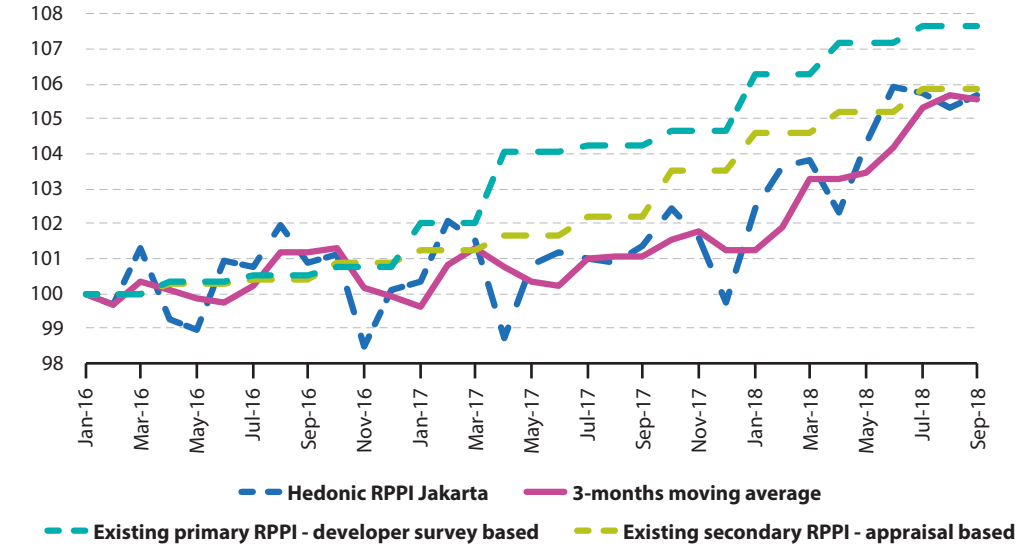
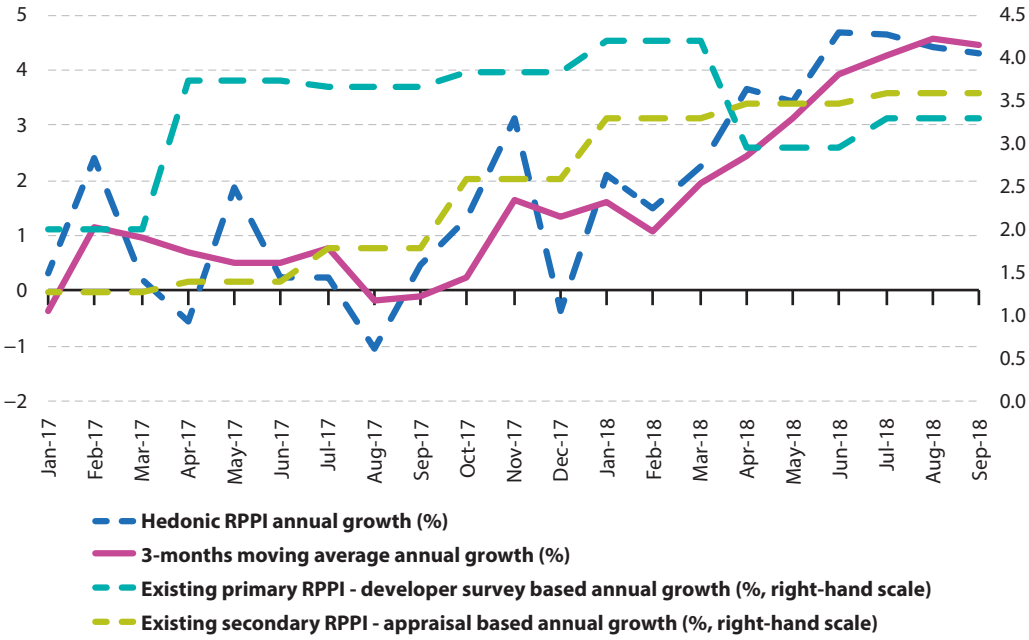



**Figure 5:** Annual growth rates for house markets, all of Jakarta, January 2017-September 2018
(%)



# 4. Conclusion

Using the time-dummy hedonic regression method we have computed alternative residential property price indices for the secondary house market in five districts of Jakarta based on property advertisements found on the web. These hedonic indices show promising results and have the potential in the future to replace the methods currently used to compile the RPPI for Indonesia. The regression outputs represent robust 'baseline' models for index compilation. Advertisement observations based on web listings seem more homogenous in nature, as indicated by the high degree of explanatory power, given the limited array of characteristic variables available. Smoothing may provide a better option for publishing an index based on these hedonic methods as it reduces short-term volatility.

For further developments, we will maintain these baseline models and extend coverage to other large cities in Indonesia. This extension will depend on the suitability of listings which may be available and the relative importance of different cities, as measured by their share in the national property market (derived from mortgage data). We need to ensure this new index remains representative of current market conditions by regularly reviewing the models' performance and updating the weights. We may also seek to enhance the models in the future by including a more granular spatial adjustment (location advantage) and other characteristics (such as the age of each property).

# Acknowledgements

The author would like to thank Niall O'Hanlon ( International Monetary Fund) for valuable technical assistance, Annisa Cynthia and Arinda Dwi Okfantia (both Bank Indonesia) for their helpful input. All views expressed are those of author and do not necessarily represent the views of the Bank Indonesia.

# References

Daas, P. J. H., M. Puts, M. Tennekes and A. Priem (2014), 'Big Data as a Data Source for Official Statistics: experiences at Statistics Netherlands', proceedings of Statistics Canada Symposium 2014, Ottawa.

Diewert, W. E. (2003), 'Hedonic Regressions. A Consumer Theory Approach', National Bureau of Economic Research in *Scanner Data and Prices Indexes*, pp. 317-348.

Eurostat (2013), *Handbook on Residential Property Prices Indices (RPPIs)*, Publications Office of the European Union, Luxembourg.

Hammer, C., D. C. Kostroch, G. Quirós, and STA Internal Group (2017), 'Big Data: Potential, Challenges and Statistical Implications', Staff Discussion Note, SDN/17/06, International Monetary Fund, Washington D.C.

Hill, R. J. (2011), 'Hedonic Price Indexes for Housing', *OECD Statistics Working Papers*, Working Paper No. 36, OECD, Paris.

Hill, R. J. (2012), 'Hedonic Price Indexes for Residential Housing: A Survey, Evaluation and Taxonomy', *Journal of Economic Surveys*, Volume 27, Issue 5, pp. 879-914.

Hülagü, T., E. Kizilkaya, A. G. Özbekler, and P. Tunar (2015), 'A Hedonic House Price Index for Turkey', presented to a Turkish Statistical Institute seminar and the European Real Estate Society 22nd Annual Conference, Istanbul.

Land Economy and Construction Industries Bureau (2016), 'Methodology of JRPPI: Japan Residential Property Price Index', Ministry of Land, Infrastructure, Transport and Tourism, Tokyo.

Li, W., M. Prud'homme and K. Yu (2006), 'Studies in Hedonic Resale Housing Price Indexes', presented to the Canadian Economic Association 40th Annual Meetings, Concordia University, Montréal.

Lyons, R. C. (2018), 'Can list prices accurately capture housing price trends? Insights from extreme markets conditions', *Finance Research Letters*, Volume 30, pp. 228-232.

Marsden, J. (2015), 'House prices in London — an economic analysis of London's housing market', *GLA Economics*, Working Paper 72, Greater London Authority.

O'Hanlon, N. (2011), 'Constructing a National House Price Index for Ireland', *Journal of the Statistical and Social Inquiry Society of Ireland*, Volume XL, pp. 167-196.

Radermacher, W. J. (2018), 'Official Statistics in the Era of Big Data Opportunities and Threats', *International Journal of Data Science and Analytics*, Volume 6, Issue 3, pp. 225-231.

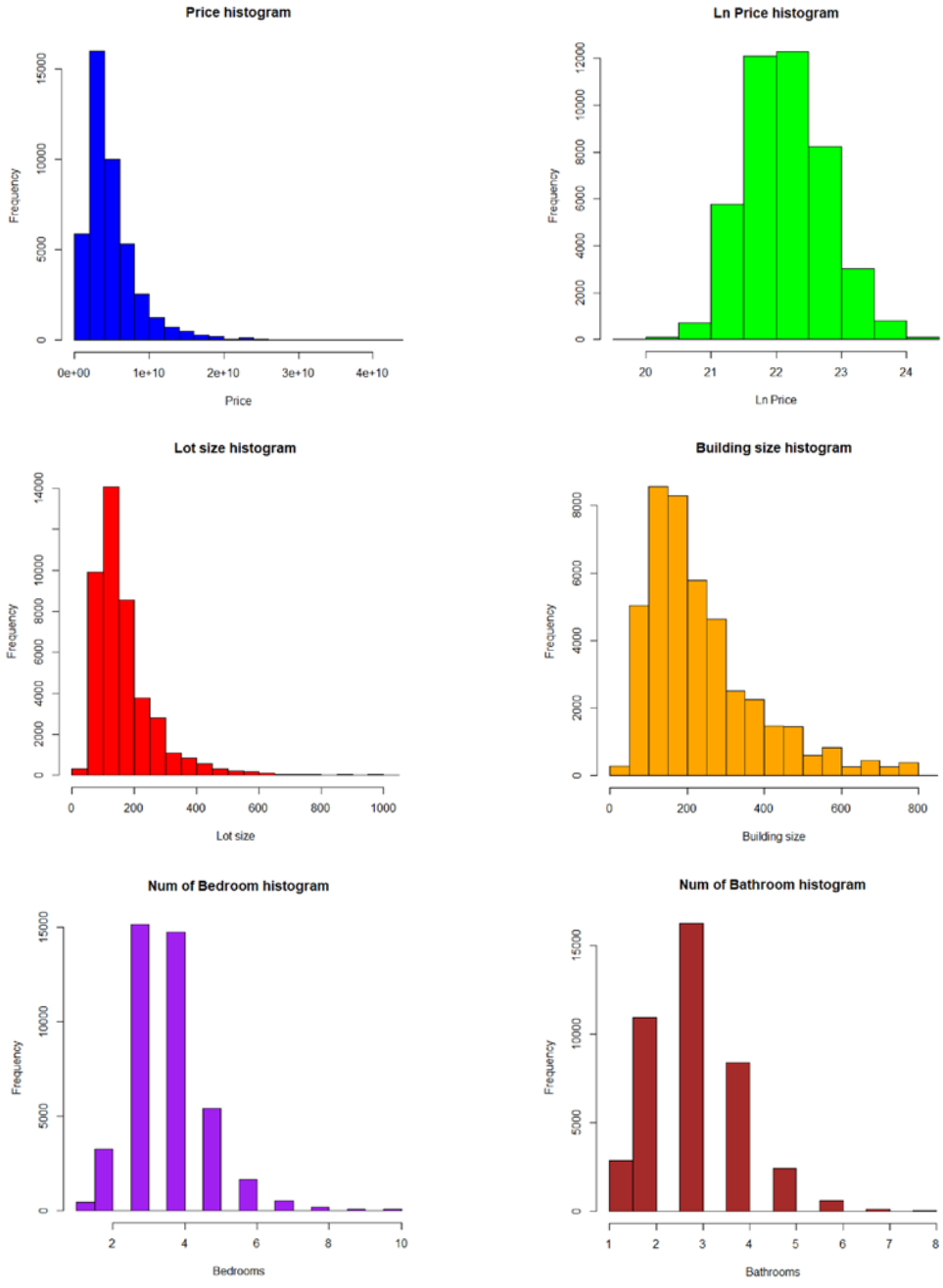
Silver, M. (2016), 'How to Better Measure Hedonic Residential Property Price Indexes', *IMF Working Papers*, WP/16/213, International Monetary Fund, Washington D.C.

Zwick, M. (2017), 'Introduction to Big Data in Official Statistics', Institute for Research and Development in Official Statistics, Federal Statistics Office Germany, Wiesbaden.

# Data appendix

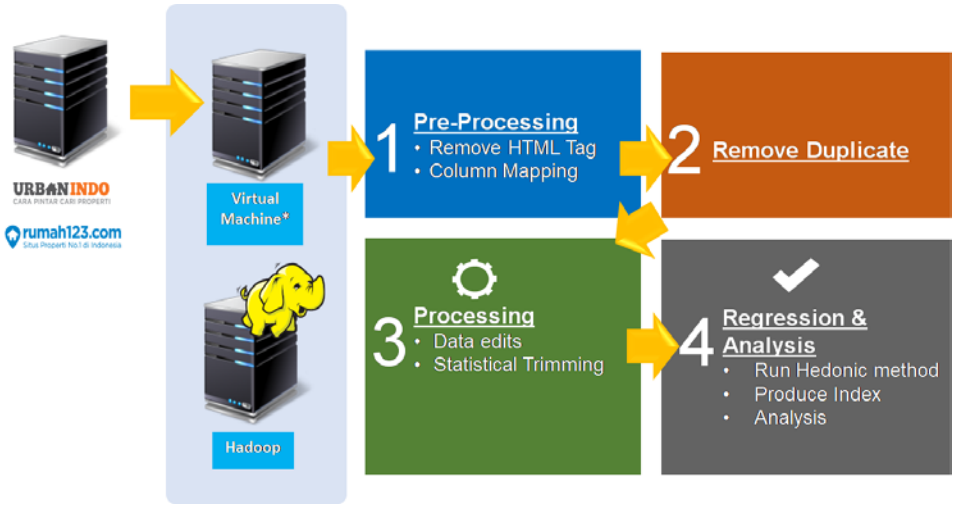
## 1. Data distribution for the sample, North Jakarta

**Figure 6:** Characteristic variables



## 2. Steps for data preparation

**Figure 7:** Workflow for the construction of an RPPI using big data




### DATA SOURCE

- Online property advertisements from the two biggest property websites with approximately 9 000 new advertisements per month (for Jakarta only).
- Data available since 2015, listings only refer to the first instance that the price of a property was listed (a unique price).
- Data attributes:
  - title;
  - status of property : sell/rent;
  - type of property (house/apartment (flat)/villa/condotel/condominium);
  - advertising time start date and end date;
  - property price;
  - land and building size;
  - number of bedrooms and number of bathrooms;
  - address;
  - property description.

### PRE-PROCESSING

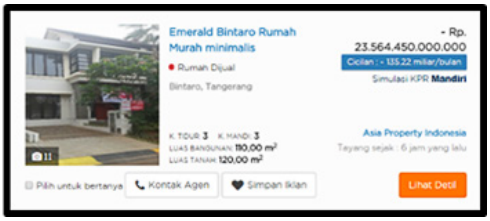
- **Cleaning**
  The cleaning process that formed part of the data pre-processing exercise included removing irrelevant characters such as HTML tags that formed part of the title and description for each advertisement. The removal of HTML tags was done through the Python programming language, using one of its libraries, HTMLParser.

- **City mapping**
  City mapping was conducted to standardise the addresses shown in the data to the city level. This process was carried out because some portals do not provide data pertaining to the city/district for each advertisement. City mapping used a list of city/districts and sub-districts obtained from Indonesian Statistics (BPS). If the address in the advertisement could not be found in the list of sub-districts, city mapping was carried out using the Geocoding API provided by Google Maps.

- **Column mapping**
  Data from different portals had distinct formats and column structures. For example, data from Rumah123 consisted of 21 attribute columns, using '|' (a bar) or '~' (a tilde) as a delimiter for the columns. In contrast, data from Urbanindo consisted of 13 attribute columns using a tab as the delimiter. Thus, column mapping was needed to standardise column structures, column names and the use of delimiters. Once completed, this allowed data from disparate sources to be compiled and processed simultaneously.

- **Removal of duplicates**
  After the initial processing, there were still many duplicate data entries. These included intra-portal and inter-portal duplicates. Intra-portal duplicates existed when the same property was advertised by different estate agents or when an advertisement was reposted by the same estate agent. Inter-portal duplications existed if the same property was advertised on different property portals. Property advertisements were considered the same if their price, land area, building area, number of bathrooms, number of bedrooms and city had the same value.

- **Removal of abnormal data**
  Aside from the issue of duplicate advertisements, there were also concerns regarding abnormal data. Below are some criteria which were used to identify abnormal data:
  a. Missing values — the advertisement did not provide data regarding the land area and/or the building's area.
  b. Unusual price — for example, land which cost Rp 0.00 or land of 100 $m^2$ which cost Rp 23 trillion.
  c. Unusual building or land area — for example, land area equal to 0 $m^2$ or a building area of 10 000 $m^2$.


  d. Unusual price due to location — for example, there was an advertisement for a house which was being sold at Rp 50 million in Jakarta Pusat.
  e. Unusual ratio of land area to building area — for example, an advertisement showed a house with a building area of 300 $m^2$ and a land area of 30 $m^2$, or a house with a building area of 30 $m^2$ and a land area of 1 000 $m^2$.

## 3. Property market share

**Table 4:** Comparisons of shares for weights and the number of observations
(%)

| Districts | Mortgage collateral share | Number of observation share |
|---|---|---|
| West Jakarta | 24.95 | 24.14 |
| Central Jakarta | 12.06 | 4.25 |
| South Jakarta | 31.92 | 37.34 |
| East Jakarta | 7.27 | 21.95 |
| North Jakarta | 23.80 | 12.31 |

## 4. Regression output

**Table 5:** Regression output — check for stability over time, sample for North Jakarta

| Dependent variable: Ln Price | | | | |
|---|---|---|---|---|
| **Independent variables** | **2016:1 - 2016:12 coefficients** | | **2017:1 - 2017:12 coefficients** | |
| Intercept | 21.2900 | *** | 21.2900 | *** |
| Building size | 0.0010 | *** | 0.0011 | *** |
| Lot size | 0.0045 | *** | 0.0041 | *** |
| Dum_# of bedroom 1-2 | -0.2153 | *** | -0.0310 | *** |
| Dum_# of bedroom 3 | -0.0440 | *** | -0.2185 | *** |
| Dum_# of bedroom >4 | -0.0400 | *** | -0.0326 | *** |
| Dum_# of bathroom 1 | -0.2225 | *** | -0.1997 | *** |
| Dum_# of bathroom 2 | -0.1732 | *** | -0.1853 | *** |
| Dum_# of bathroom >3 | 0.0082 | * | 0.0038 | |
| Dum_period 2016 (2017):2 | 0.0007 | | 0.0147 | |
| Dum_period 2016 (2017):3 | 0.0006 | | 0.0025 | |
| Dum_period 2016 (2017):4 | 0.0034 | | -0.0077 | |
| Dum_period 2016 (2017):5 | -0.0203 | *** | -0.0012 | |
| Dum_period 2016 (2017):6 | -0.0123 | | -0.0049 | |
| Dum_period 2016 (2017):7 | -0.0132 | | 0.0167 | * |
| Dum_period 2016 (2017):8 | 0.0022 | | 0.0244 | *** |
| Dum_period 2016 (2017):9 | -0.0143 | | -0.0061 | |
| Dum_period 2016 (2017):10 | -0.0169 | ** | 0.0009 | |
| Dum_period 2016 (2017):11 | -0.0307 | *** | -0.0023 | |
| Dum_period 2016 (2017):12 | -0.0435 | *** | -0.0195 | * |
| Adjusted R-squared | 0.863 | | 0.862 | |
| F-statistics | 4 866 | | 4 067 | |
| Number of observations | 14 640 | | 14 026 | |

Note: *** significance at 1 %, ** significance at 5 % and * significance at 10 %.

**Table 6:** Testing for heteroscedasticity

| Breusch-Pagan test for heteroscedasticity | |
|---|---|
| North Jakarta | BP = 204.1, df = 19, p-value < 2.2e-16 |
| South Jakarta | BP = 366.96, df = 19, p-value < 2.2e-16 |
| Central Jakarta | BP = 109.83, df = 19, p-value = 8.574e-15 |
| West Jakarta | BP = 412.43, df = 19, p-value < 2.2e-16 |
| East Jakarta | BP = 135.36, df = 19, p-value < 2.2e-16 |

**Table 7:** Regression output for three other districts of Jakarta

| | Central Jakarta | | West Jakarta | | East Jakarta | |
|---|---|---|---|---|---|---|
| **Dependent variable: Ln Price** | | | | | | |
| **Independent variables** | **Estimates** | **Robust standard error** | **Estimates** | **Robust standard error** | **Estimates** | **Robust standard error** |
| Intercept | 21.0800 | 0.02999 *** | 20.9310 | 0.00682*** | 20.5700 | 0.00906*** |
| Building size | 0.0014 | 0.00006 *** | 0.0012 | 0.00002*** | 0.0013 | 0.00003*** |
| Lot size | 0.0038 | 0.00007 *** | 0.0044 | 0.00003*** | 0.0033 | 0.00003*** |
| Dum_# of bedroom 1-2 | -0.0604 | 0.01616 *** | -0.0382 | 0.00351*** | -0.0666 | 0.00482*** |
| Dum_# of bedroom 3 | -0.1352 | 0.02586 *** | -0.2171 | 0.00525*** | -0.2564 | 0.00773*** |
| Dum_# of bedroom >4 | -0.1415 | 0.01474 *** | -0.0375 | 0.00483*** | -0.0250 | 0.00590*** |
| Dum_# of bathroom 1 | -0.4678 | 0.02789 *** | -0.0968 | 0.00597*** | -0.3393 | 0.00799*** |
| Dum_# of bathroom 2 | -0.1743 | 0.01564 *** | -0.0974 | 0.00336*** | -0.1387 | 0.00456*** |
| Dum_# of bathroom >3 | 0.0084 | 0.01505 | 0.0194 | 0.00436*** | 0.0293 | 0.00534*** |
| Dum_period 2016:2 | 0.0185 | 0.03000 | -0.0096 | 0.00676 | 0.0001 | 0.00895 |
| Dum_period 2016:3 | 0.0752 | 0.02998 ** | -0.0191 | 0.00667*** | 0.0414 | 0.00890*** |
| Dum_period 2016:4 | 0.0275 | 0.02962 | -0.0253 | 0.00668*** | 0.0064 | 0.00898 |
| Dum_period 2016:5 | -0.0306 | 0.02658 | -0.0295 | 0.00578*** | 0.0326 | 0.00779*** |
| Dum_period 2016:6 | 0.0164 | 0.03036 | -0.0202 | 0.00706*** | 0.0383 | 0.00947*** |
| Dum_period 2016:7 | 0.0816 | 0.03128 *** | -0.0103 | 0.00703 | 0.0202 | 0.00944** |
| Dum_period 2016:8 | 0.0966 | 0.03310 *** | -0.0146 | 0.00693** | 0.0590 | 0.00910*** |
| Dum_period 2016:9 | 0.0503 | 0.03384 | -0.0405 | 0.00741*** | 0.0705 | 0.00950*** |
| Dum_period 2016:10 | 0.0577 | 0.02918 ** | -0.0164 | 0.00656** | 0.0396 | 0.00839*** |
| Dum_period 2016:11 | 0.0066 | 0.03123 | -0.0259 | 0.00687*** | 0.0320 | 0.00890*** |
| Dum_period 2016:12 | 0.0766 | 0.03284 ** | -0.0098 | 0.00742 | 0.0606 | 0.00942*** |
| Adjusted R-squared | 0.731 | | 0.813 | | 0.835 | |
| F-statistics | 667 | | 7 138 | | 5 891 | |
| Number of observations | 4 662 | | 31 184 | | 22 180 | |

Note: *** significance at 1 %, ** significance at 5 % and * significance at 10 %.

## 5. Results for hedonic RPPI

**Figure 8:** Comparison of indices for the secondary market for houses, West Jakarta, January 2016-September 2018
(January 2016 = 100)

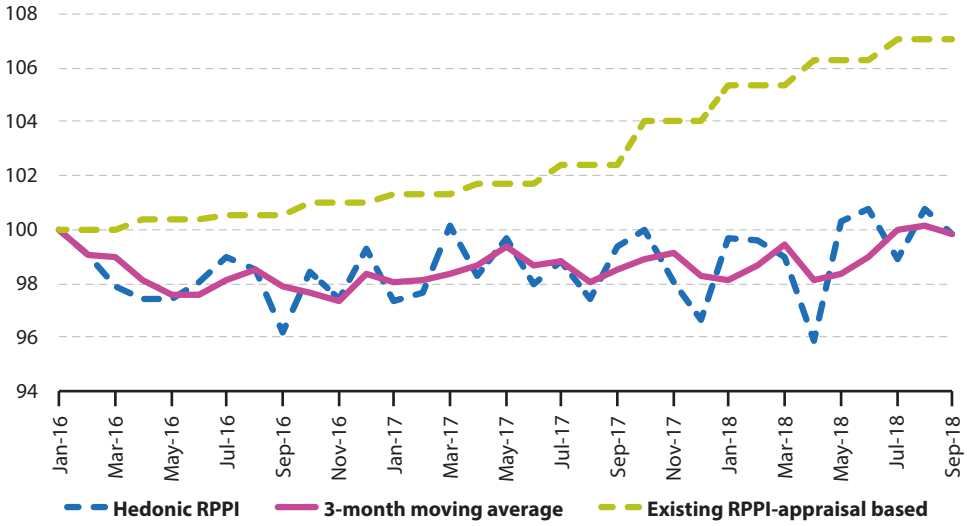



**Figure 9:** Comparison of indices for the secondary market for houses, East Jakarta, January 2016-September 2018
(January 2016 = 100)

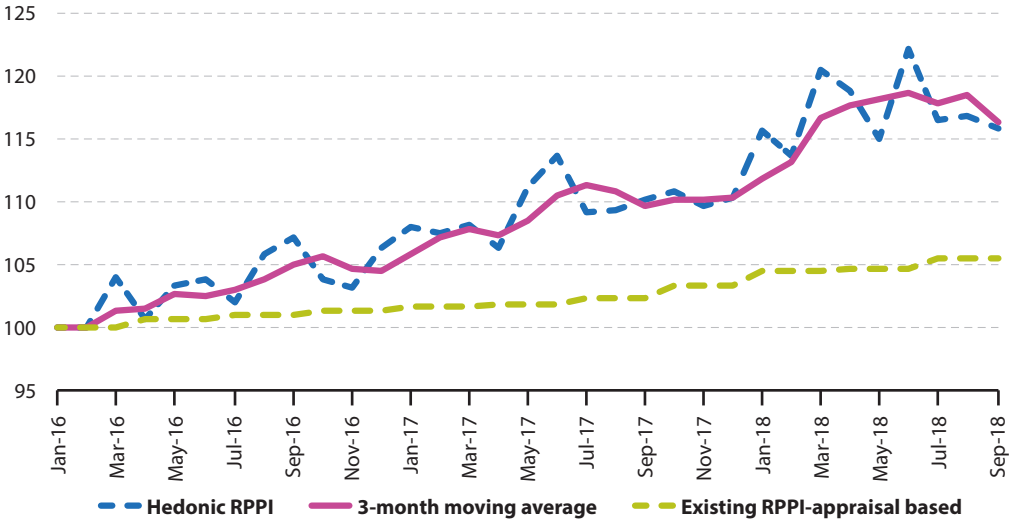


**Table 8:** Hedonic RPPI, January 2016-September 2018
(January 2016 = 100)

| Period | West Jakarta | | Central Jakarta | | South Jakarta | | East Jakarta | | North Jakarta | | All of Jakarta | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TD Hedonic | Adjusted* | TD Hedonic | Adjusted* | TD Hedonic | Adjusted* | TD Hedonic | Adjusted* | TD Hedonic | Adjusted* | TD Hedonic | Adjusted* |
| Jan-16 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Feb-16 | 99.07 | 99.07 | 103.33 | 103.33 | 98.53 | 98.53 | 100.01 | 100.01 | 100.02 | 100.02 | 99.71 | 99.71 |
| Mar-16 | 97.92 | 99.00 | 108.87 | 104.07 | 101.47 | 100.00 | 104.09 | 101.37 | 99.98 | 100.00 | 101.31 | 100.34 |
| Apr-16 | 97.43 | 98.14 | 102.98 | 105.06 | 98.24 | 99.41 | 100.67 | 101.59 | 100.23 | 100.08 | 99.26 | 100.09 |
| May-16 | 97.39 | 97.58 | 98.56 | 103.47 | 100.00 | 99.90 | 103.37 | 102.71 | 98.18 | 99.46 | 98.98 | 99.85 |
| Jun-16 | 98.03 | 97.61 | 101.55 | 101.03 | 103.78 | 100.67 | 103.82 | 102.62 | 98.94 | 99.12 | 100.93 | 99.72 |
| Jul-16 | 98.95 | 98.12 | 109.84 | 103.32 | 100.17 | 101.32 | 102.06 | 103.08 | 98.50 | 98.54 | 100.77 | 100.23 |
| Aug-16 | 98.52 | 98.50 | 111.52 | 107.64 | 101.40 | 101.78 | 105.95 | 103.94 | 100.15 | 99.20 | 101.93 | 101.21 |
| Sep-16 | 96.16 | 97.88 | 105.72 | 109.03 | 103.02 | 101.53 | 107.23 | 105.08 | 98.49 | 99.05 | 100.86 | 101.19 |
| Oct-16 | 98.40 | 97.69 | 106.69 | 107.98 | 102.55 | 102.33 | 103.84 | 105.67 | 98.32 | 98.98 | 101.10 | 101.30 |
| Nov-16 | 97.46 | 97.34 | 101.67 | 104.69 | 98.28 | 101.29 | 103.28 | 104.78 | 96.80 | 97.87 | 98.49 | 100.15 |
| Dec-16 | 99.27 | 98.37 | 107.74 | 105.37 | 99.68 | 100.17 | 106.35 | 104.49 | 95.78 | 96.97 | 100.10 | 99.90 |
| Jan-17 | 97.38 | 98.03 | 110.65 | 106.69 | 99.86 | 99.27 | 108.08 | 105.90 | 96.42 | 96.34 | 100.32 | 99.64 |
| Feb-17 | 97.64 | 98.09 | 116.20 | 111.53 | 101.90 | 100.48 | 107.50 | 107.31 | 98.25 | 96.82 | 102.10 | 100.84 |
| Mar-17 | 100.10 | 98.37 | 115.45 | 114.10 | 99.54 | 100.44 | 108.22 | 107.93 | 96.56 | 97.08 | 101.52 | 101.31 |
| Apr-17 | 98.29 | 98.68 | 107.02 | 112.89 | 96.14 | 99.19 | 106.35 | 107.35 | 96.13 | 96.98 | 98.73 | 100.78 |
| May-17 | 99.67 | 99.36 | 109.17 | 110.54 | 98.78 | 98.15 | 111.30 | 108.62 | 97.35 | 96.68 | 100.83 | 100.36 |
| Jun-17 | 98.00 | 98.65 | 109.69 | 108.63 | 101.21 | 98.71 | 113.79 | 110.48 | 96.39 | 96.62 | 101.20 | 100.25 |
| Jul-17 | 98.79 | 98.82 | 108.11 | 108.99 | 100.15 | 100.05 | 109.28 | 111.46 | 98.41 | 97.38 | 101.02 | 101.01 |
| Aug-17 | 97.41 | 98.06 | 107.22 | 108.34 | 100.52 | 100.63 | 109.46 | 110.84 | 99.18 | 97.99 | 100.88 | 101.03 |
| Sep-17 | 99.38 | 98.53 | 117.33 | 110.89 | 98.77 | 99.81 | 110.20 | 109.65 | 96.07 | 97.89 | 101.35 | 101.08 |
| Oct-17 | 99.97 | 98.92 | 112.03 | 112.19 | 103.03 | 100.77 | 110.91 | 110.19 | 96.75 | 97.33 | 102.43 | 101.55 |
| Nov-17 | 98.08 | 99.14 | 110.64 | 113.33 | 102.84 | 101.55 | 109.77 | 110.29 | 96.51 | 96.44 | 101.59 | 101.79 |
| Dec-17 | 96.68 | 98.24 | 105.01 | 109.22 | 101.27 | 102.38 | 110.39 | 110.36 | 95.04 | 96.10 | 99.76 | 101.26 |
| Jan-18 | 99.67 | 98.14 | 102.60 | 106.08 | 104.48 | 102.86 | 115.69 | 111.95 | 98.38 | 96.64 | 102.41 | 101.25 |
| Feb-18 | 99.60 | 98.65 | 110.33 | 105.98 | 106.74 | 104.16 | 113.74 | 113.27 | 97.16 | 96.86 | 103.62 | 101.93 |
| Mar-18 | 98.99 | 99.42 | 110.92 | 107.95 | 106.77 | 106.00 | 120.51 | 116.65 | 96.25 | 97.26 | 103.82 | 103.29 |
| Apr-18 | 95.84 | 98.14 | 103.53 | 108.26 | 106.28 | 106.60 | 118.85 | 117.70 | 98.25 | 97.22 | 102.35 | 103.26 |
| May-18 | 100.31 | 98.38 | 110.13 | 108.19 | 107.55 | 106.87 | 115.07 | 118.14 | 97.76 | 97.42 | 104.27 | 103.48 |
| Jun-18 | 100.73 | 98.96 | 110.51 | 108.06 | 111.63 | 108.49 | 122.19 | 118.70 | 96.46 | 97.49 | 105.93 | 104.18 |
| Jul-18 | 98.93 | 99.99 | 113.70 | 111.45 | 112.50 | 110.56 | 116.50 | 117.92 | 96.35 | 96.86 | 105.71 | 105.30 |
| Aug-18 | 100.73 | 100.13 | 111.81 | 112.01 | 108.74 | 110.96 | 116.93 | 118.54 | 98.79 | 97.20 | 105.34 | 105.66 |
| Sep-18 | 99.84 | 99.83 | 112.41 | 112.64 | 112.34 | 111.20 | 115.91 | 116.45 | 96.38 | 97.18 | 105.69 | 105.58 |

*Adjusted using 3-months moving average.

# 4 Measuring price dynamics of package holidays with transaction data [1]

KAROLA HENN [2], CHRIS-GABRIEL ISLAM [2], PATRICK SCHWIND [3] AND ELISABETH WIELAND [4]

**Abstract**: In Germany, package holidays, which consist of a bundle of flight and accommodation services, are an important driver of consumer prices. Several challenges arise when measuring the price development of package holidays, for example the quality of accommodation, the timing of the booking, the treatment of out-of-season services as well as the underlying holiday season. Statistical practices are currently based on sampling offer prices. As a possible alternative, transaction price data from a commercial booking system are analysed in this study. Our dataset comprises both online bookings and bookings made via stationary travel agencies of package holidays on a daily basis. The large sample size allows for a disaggregation by individual holiday destination.

The paper analyses the possibilities and challenges in compiling a price index out of transaction data for flight package holidays. The dataset raises a number of methodological issues, for example the grouping of unstructured text information into meaningful categories, the handling of missing information or the identification of outliers. Moreover, various index aggregation methods are analysed, which include hedonic regressions, stratification, and also a multilateral index method. Applied to six major holiday destinations for German travellers, all transaction-based methods under consideration exhibit similar price dynamics, pointing to robust results for destination-based price indicators for package holidays.

**JEL codes**: C14, C43, E31

**Keywords**: consumer prices, transaction data, hedonic regressions, quality adjustment, multilateral index number methods

(1) This paper represents the authors' personal opinions and does not necessarily reflect the views of Destatis, the Deutsche Bundesbank, the European Central Bank or the Eurosystem.

(2) Federal Statistical Office (Destatis), Prices, Methods and Communication of Price Statistics Division.
(3) Deutsche Bundesbank, Directorate Statistics, General Economic Statistics Division.
(4) Deutsche Bundesbank, Directorate Economics, Macroeconomic Analysis and Projections Division.

# 1. Motivation

In traditional price collection, offer prices from pre-defined price representatives are usually collected at fixed points in time every month. The more complex a given good or service is, the more manual work is required by a national statistical institute (NSI) in setting-up a sufficiently large selection of price representatives. This is especially true for bundles of different services, such as package holidays, which are made up of both travel and accommodation (hotel) services and have a lot of price-determining characteristics such as the category of the hotel as well as the meal type, the room or the departure airport. Moreover, travel-related prices such as the flight can fluctuate heavily within a given month.

In German price statistics, package holidays have a weight of 2.7 % in the Harmonised Index of Consumer Prices (HICP) as of 2019. However, due to their high volatility and strong seasonality, package holidays have a noticeable effect on the German and even the euro area inflation rate. The Federal Statistical Office (Destatis) currently uses a global distribution system from information technology (IT) provider Amadeus (Amadeus Germany GmbH) — as applied by travel agencies — to collect offer prices of package holidays. The sample size is limited due to the high effort required for manual price collection. Therefore, it is currently not possible to publish price developments broken down by holiday destinations, rather, broad subindices are published for 'Domestic package holidays' (ECOICOP 09.6.0.1) and 'International package holidays' (ECOICOP 09.6.0.2) ([5]).

An alternative to collecting offer prices consists in transaction data ([6]) by using actual bookings of international package holidays recorded in the Amadeus IT booking systems, which are used by online travel agencies or at traditional high street travel agencies. The aim of this paper is to investigate the possibilities and challenges when compiling a price index out of transaction data for flight package holidays ([7]), which are very heterogeneous seasonal services ([8]). Due to the large sample size of the underlying transaction dataset, the resulting experimental price index could be subdivided into relevant holiday destinations, thus allowing for a more detailed economic interpretation of the underlying price movements of package holidays. In particular, such destination-based price indicators could help to disentangle the overall price trend in package holidays from short-term movements for a given holiday destination, which would provide a high level of value added for consumer price analysis. This paper also contributes through the application of the most recent index aggregation methods, which include hedonic regressions, stratification, and a multilateral method, to the relatively new field of measuring prices of (bundled) services by transaction data.

([5]) The goods and services in the HICP follow the European Classification of Individual Consumption according to Purpose (ECOICOP). For an overview of this classification, see: https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL&StrNom=COICOP_5&StrLanguageCode=EN&IntPcKey=&StrLayoutCode=HIERARCHIC&IntCurrentPage=1.

([6]) Note that transaction prices are generally in line with the basic price definition in the HICP: 'The prices used in the HICP should be purchaser prices, which are the prices actually paid by households' (see Eurostat, 2018, p. 30).

([7]) Besides flight package holidays, the German HICP subindex for package holidays also consists of domestic package holidays, shorter city trips to other European countries and cruises (see Section 2), which were not the subject of this study.

([8]) Although interesting on its own, this paper does not analyse the seasonality of package holidays itself such as the imputation of out-of-season package holidays.

The paper is structured as follows: Section 2 describes the current official practice in measuring prices of package holidays by the Federal Statistical Office, which is based on offer prices. Section 3 presents the transaction dataset from Amadeus and comments on the challenges of processing these data for the purpose of price statistics. Section 4 discusses various methods commonly used to measure prices, as well as newer index methods that have recently been developed on the basis of scanner data. Section 5 compares the price indices derived from the various methods for six major holiday destinations of German travellers. Section 6 concludes and provides an outlook on the feasibility of destination-based price indicators for package holidays.

## 2. Current official practice for the German HICP

In official price statistics, package holidays reflect a bundled cost of travel and accommodation services sold in one transaction, for example a return flight in combination with a seven day hotel stay. By convention, the price of a package holiday enters the official HICP always in the month during which the holiday takes place and not in the month during which the holiday is booked (see Eurostat (2018), Chapter 12.5). Nevertheless, the timing of when the booking was made (for example early or last minute bookings) is an important price determinant of a package holiday. Thus, official price statistics typically use booking prices from different points in time ahead when compiling a price index for a given travel month.

To calculate the official HICP subindex for package holidays, the German Federal Statistical Office collects offer prices. This data represent a very detailed specified sample of trips, with the aim of ensuring a pure price comparison. According to the EU regulation, two methods are allowed for calculating indices for package holidays: the fixed weights method (also known as strict annual weights) and a class-confined seasonal weights method ([9]). Before the German national CPI was revised and rebased to 2015 = 100 in February 2019, the class-confined seasonal weights method was used, with a different summer and winter sample. From reporting year 2015 onwards, the official HICP subindex for package holidays is based on the fixed weights method, where the missing prices for out-of-season months are imputed ([10]).

([9]) See European Commission Regulation No 330/2009, Article 2, as well as Eurostat (2018), Chapter 7.1 on seasonal products and Chapter 12.5 on flights and package holidays.

([10]) Switching to CPI basis 2015 and using the fixed weights methods improved the interpretability of the previous month's rate of change in April, May and November of a year. At the same time, it increased the seasonal profile of the package holiday price index, with higher index values in the summer and lower values in the winter season. See also Eurostat (2019) and Deutsche Bundesbank (2019).

Table 1 provides an overview of the elementary aggregates of the German HICP for package holidays (ECOICOP 09.6). The sample for the subindex for 'international package holidays' consists of holidays from Germany to six holiday destinations (the Balearic Islands, the Canary Islands, Greece, Turkey, Egypt and the Dominican Republic) with a duration of 7-14 days and to two countries for shorter city trips. Moreover, the international aggregate includes cruises. For most holiday destinations, there exist three strata: summer, winter, and whole-year strata. Missing prices for the summer sample for a given holiday destination are imputed using the winter or the whole-year sample and vice versa (counter-seasonal estimation). For two holiday destinations, there is only a summer or winter sample and missing prices are imputed using all other available prices (all-seasonal estimation).

**Table 1: Elementary aggregates for the German HICP subindex for package holidays**

| ECOICOP | Weight in total package holidays (%) | Coverage | Sample period |
|---|---|---|---|
| **09.6.0.1 Domestic package holidays** | 5.60 | Germany only, travel by train or car | Summer/winter |
| **09.6.0.2 International package holidays** | | | |
| International flight package holidays (7 to 14 days) | 76.95 | Four holiday destinations | Summer/winter/whole year |
| | | Two holiday destinations | Summer or winter only |
| City trips | | Two holiday destinations | Whole year |
| Cruises | 17.45 | Combination of flight and open-sea cruise | Summer only |

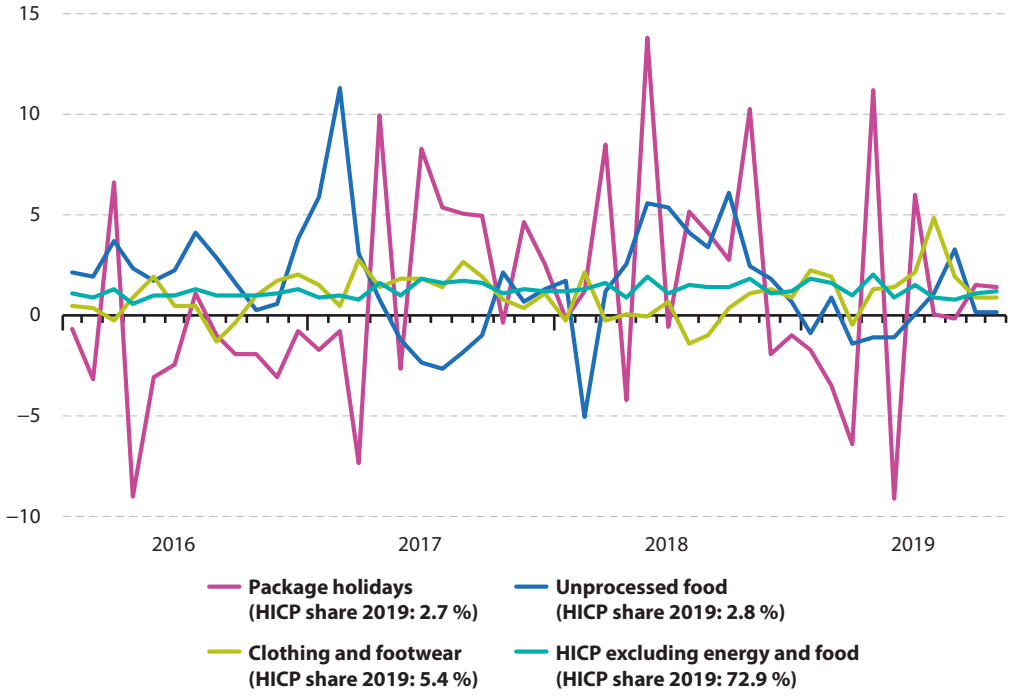
Note: as covered by ECOICOP 09.6.

In German price statistics, offer prices for international package holidays are collected from the booking system *START Amadeus* ([11]) via the internet and cover roughly 300 price representatives. Booking codes from tour operators are used to identify a product offer with pre-defined attributes (for example hotel XXXX, all inclusive, double room with sea view, for two persons and ten days, with departure flight from Frankfurt am Main). The price representatives are calculated using three offer prices (three inquiries at different points in time in advance of a given departure) for the winter/summer sample or 21 offer prices (three inquiries in advance of seven departure days) for the whole-year sample. In total, about 1 500 to 3 000 offer prices (depending on the timing of public holidays) are included in the price calculation for a given travel month.

([11]) The booking system *START Amadeus* is used by traditional high street travel agencies to handle booking transactions for package holidays (see Section 3 for more information on the data provider). In contrast, the offer prices for city trips are collected manually from different online travel agencies, whereas for cruises, catalogue prices are compiled.

The resulting German HICP subindex for package holidays exhibits a high degree of volatility, as shown in Figure 1. The annual rate of change from January 2016 onwards ranges between -9 and +14 percentage points and is therefore more volatile than other seasonal HICP components, such as clothes or unprocessed food ([12]). From the perspective of a data user, a more detailed breakdown by holiday destinations would be helpful in interpreting such price movements ([13]). From an international perspective, the weight of package holidays in the German HICP (2019: 2.7 %) is one of the highest among European countries, with higher values only observed in Iceland (6.3 %), the United Kingdom (4.2 %) and Cyprus (3.2 %). Because of its weight and volatility, the challenges of measuring prices for package holidays with transaction data and how to derive prices for bundled services, which are generally more complex than supermarket goods, are very important to Germany, but may be relevant to other (European) NSIs as well ([14]).

**Figure 1: German HICP package holidays compared with other components**
(annual rate of change, %)



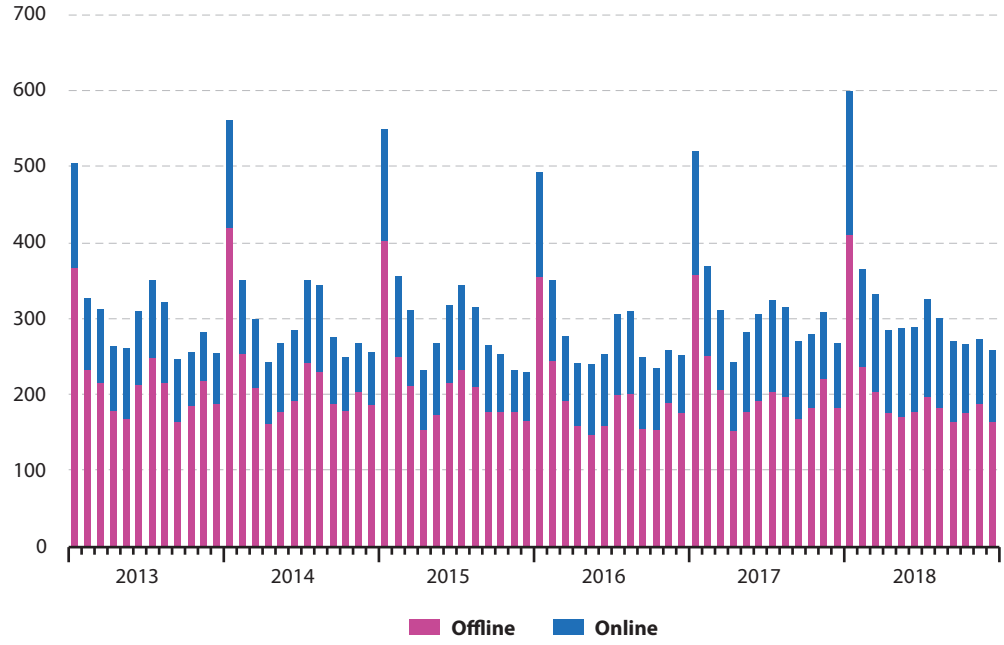

*Source:* Eurostat

([12]) Amongst others, possible contributors are Easter and/or the Pentecost holidays, which vary from year to year (unlike Christmas).
([13]) See also Deutsche Bundesbank (2017) for a comment on the impact of HICP package holidays on core inflation in Germany.
([14]) To the best of our knowledge, only the Dutch and Swedish NSIs have already implemented a transaction-based price index for package holidays in their regular index production (see, for instance, Johansson and Tongur (2019)). Both NSIs use a method that is similar to the *traditional stratification* method in this paper (see Section 4.3.2).

# 3. Description of the Amadeus dataset

The Amadeus IT Group operates an IT system for sales and marketing in the field of travelling. The underlying dataset for Germany contains around 3.7 million transaction prices per year for flight package holidays of German travellers in the period from 2013 to 2018. The data are collected via the Amadeus booking system, which is used by online travel portals as well as traditional high street travel agencies in Germany ([15]). For each transaction, information on price determinants such as the accommodation, holiday destination and number of travellers is given ([16]). The data are made up of both online and offline (in other words, via traditional high street travel agencies) bookings. The offline data constitute the larger component (see Figure 2) and usually contain two to three times as many observations as online data, but they do not contain detailed information on meal types, room categories, car rentals or travel insurance. Given the different levels of information provided as well as the possibility of different pricing methods, it may make sense to examine the online and offline booking channels separately when measuring prices.

**Figure 2: Number of offline and online transactions per booking month**
(thousand)




*Source:* Bundesbank calculations on the basis of booking data from Amadeus Leisure IT GmbH

([15]) According to the economic newspaper *WirtschaftsWoche* (issue 27/2018), Amadeus has a global market share of 43 %. See Nagengast, Bursian and Menz (2019) for an application of the Amadeus dataset in analysing the role of dynamic pricing for exchange rate pass through.
([16]) For an overview of variables from the data provider, see Table A.1 in the Appendix. Table A.2 lists the additional variables created for this paper.

Datasets that have not been compiled primarily for the purpose of price statistics may exhibit a multitude of irregularities. The transaction dataset may, for instance, be incomplete or contain incorrect entries. For example, in about 10 % of offline bookings, the holiday destination is missing. There are also cases in which the travel date (***travelDate***) is earlier than the booking date (***transactionDate***). Incorrect entries of this kind are filtered out beforehand ([17]). Moreover, outliers in the Amadeus dataset concerning the price and the duration of the package holiday are also excluded. Corresponding to the first and 99th percentile of transactions, outliers for prices per person per day are defined as those under EUR 27 or those over EUR 427 and outliers concerning the duration of the package holiday as those less than two days or more than 22 days. Overall, after adjusting for outliers, roughly 3.4 million observations per year remain for holidays in the period from 2013 to 2018.

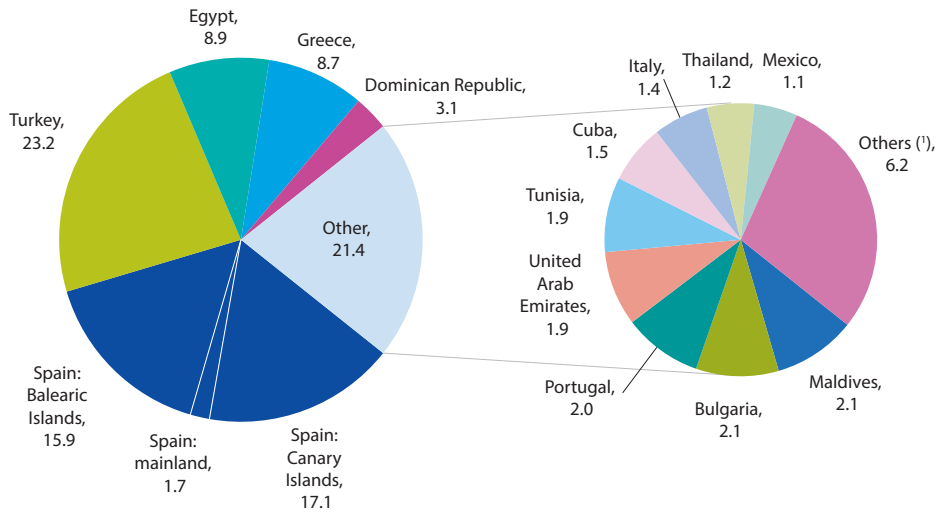
In addition to data cleansing and outlier adjustment, it is also necessary to categorise the unstructured text information in some variables of the (more detailed) online bookings. For example, more than 100 different variations exist for the online variable ***mealType***. Across the entire dataset, the number of different variations for the variable ***roomCategory*** is even higher, at 80 000. In order to categorise this level of variety, it is necessary to use string matching techniques like substring searches where the categories are defined manually in advance ([18]). Identifying children's prices, for which no set definition exists across all tour operators, represents another challenge. While offline bookings contain information on whether children are part of the booking, and if so, how many (***childrenCount***), for online bookings an assumption must be made based on the reported ages of the travellers (***travellersAges***). In the following, children were defined as travellers less than 16 years of age.

Measured by total revenue in 2015 (and excluding cruises), the most popular destinations for German travellers were Turkey (23.2 %), the Canary Islands (17.1 %), the Balearic Islands (15.9 %), Egypt (8.9 %), Greece (8.7 %) and the Dominican Republic (3.1 %), as shown in Figure 3. These six destinations together account for more than three quarters of the total revenue for German package holidays. For a disaggregation of price dynamics by destination, it therefore makes sense to focus exclusively on these six destinations ([19]). The revenue shares of the nine next most visited destinations were less than 2 % and all had fairly similar shares (range: 1.1 percentage points) ([20]). Using the transaction dataset, it is possible to derive a set of stylised facts for the German travel market. Based on data for 2015, the typical package holidaymaker travels with one other person (64 %) and without children (80 %), flies from Düsseldorf (16 %), Frankfurt (14 %) or Munich (11 %), stays for 7 or 14 days (35 % and 19 %, respectively) in a four-star hotel (59 %), and pays an average of EUR 92 per person per day.

A peculiarity of the HICP for package holidays is that bookings can, in principle, be made up to a year before departure and the timing of a booking can have an impact on the price. For the period under review, Figure 4 shows that over 20 % of all bookings had already been made half a year prior to the month of travel. On average, half of the bookings had already been made three months or more in advance. The price per person per day is 3 % more expensive than average for those holidaymakers who make their booking 6 or 12 months before departure, whereas the price falls sharply if the booking is made within two months of the departure date.

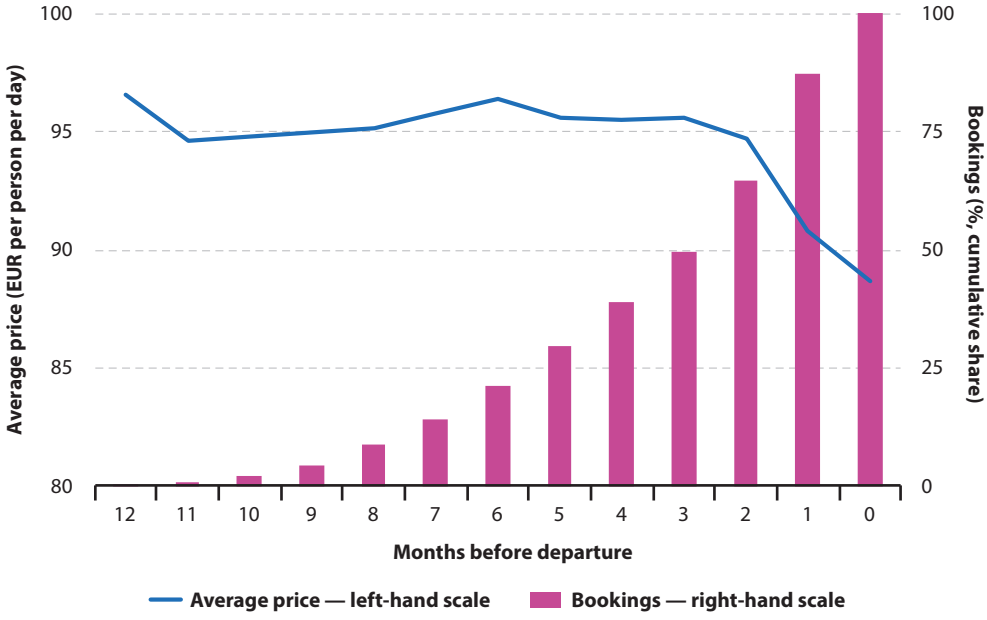
([17]) Cancellations, which are available for offline bookings only, are not included in the analysis.
([18]) See Table A.3 for the categorisation of the variable ***roomType***, which follows a kind of 'dictionary'. For the production of statistical data, this dictionary would need to be updated from time to time.
([19]) In the following, these six holiday destinations form the variable ***topArea***.
([20]) Note that the share of total revenue attributed to the six principal destinations shifted considerably over the observed period up to 2018. For example, Turkey's share fell by over half from 2013 to 2017, whereas the share of bookings for Greece and the Dominican Republic rose by roughly the same factor. In 2018, Turkey's share recovered, whereas the share of bookings for the Dominican Republic returned to its level for 2013.

**Figure 3:** Revenue shares of package holidays by destination, 2015
(%)




Note: excluding cruises.
(1) Holiday destinations with a transaction weight of less than 1 %.
*Source:* Bundesbank calculations on the basis of booking data from Amadeus Leisure IT GmbH

**Figure 4:** Bookings and average price by number of months before departure, 2013-2018
(EUR and %)



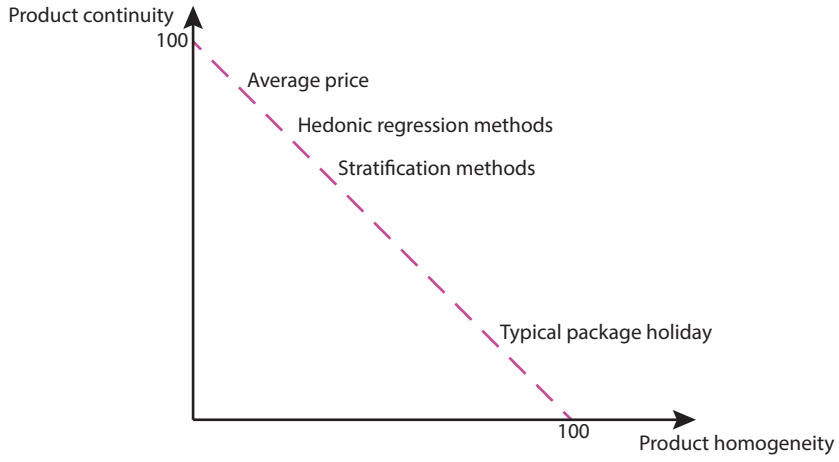

Note: excluding cruises; the left-hand axis is cut.
*Source:* Bundesbank calculations on the basis of booking data from Amadeus Leisure IT GmbH

# 4. Methods of price measurement

An ideal price index would be based on a basket of goods which compares prices of exactly the same product over time. However, transaction price data typically lack a prior product mapping, leaving it to the price statistician to define similar products within a given dataset. This process can be considered in terms of two dimensions for 'product continuity' and 'product homogeneity', when comparing transactions between any two periods. In the context of package holidays, two extrema are at hand (see Figure 5). A simple *average price* across all bookings would have the highest *product continuity*, in other words a high share of observations used over time. Still, it might be heavily affected by compositional changes in the underlying bookings, and therefore not provide a high degree of *product homogeneity* in terms of comparing similar package holidays. In contrast, the price statistician could only select transactions which correspond to a (pre-defined) *typical package holiday*. This approach coincides basically with the current official practice of collecting prices only for a given price representative. When applied to transaction data, it however does not provide a high enough number of bookings used over time to have a sound basis for any disaggregation by destination.

In the following, Section 4.1 will illustrate an average price with a high continuity of bookings, but a low degree of product homogeneity. Consequently, two main approaches in constructing a transaction-based price index are considered, both with the aim of achieving a balance between product continuity and homogeneity (see Figure 5). The first class of models will be based on *hedonic regression methods*, which estimate a price or index value by controlling for price-determining characteristics (Section 4.2). The second class of models is based on increasing the homogeneity of the bookings used by employing *stratification methods* (Section 4.3).

**Figure 5: Hypothetical trade-off between continuity and homogeneity**
(%)




Note: product continuity refers to the degree of product match in terms of observations, when comparing a given month with a base period; product homogeneity refers to the degree of similarity of items within a given product.

*Source:* Bundesbank

## 4.1. Unit value price index

The simplest approach to construct a price index is a *unit value price index*, which basically compares average prices over time. In the context of package holidays, the price per person per day (***PPD***) as given by the variables ***travellerCount*** and ***duration*** is computed for each transaction. Consequently, the average ***PPD*** for a given holiday destination is defined by:

$$\overline{PPD}_t = \frac{1}{N_t} \cdot \sum_{i=1}^{N_t} \frac{totalPrice_{i,t}}{travellerCount_{i,t}} \Big/ duration_{i,t}, \quad (1)$$

where $i = 1, \ldots, N_t$ denotes the number of transactions in period $t$ ([21]). For comparison purposes, the series of average prices are rebased to 2015=100. The resulting *unit value price index*, $I_t^{UV}$, in period $t$ is given by:

$$I_t^{UV} = \frac{\overline{PPD}_t}{\overline{PPD}_{2015}} \cdot 100. \quad (2)$$

The *unit value price index* is often applied in the context of export and import price indices and is suitable for aggregating identical, homogeneous products such as fuel and electricity (see, for instance, IMF (2009)). However, for more complex or heterogeneous products, this index would suffer from a *unit value bias* related to compositional changes in the underlying basket of goods. An example for this bias consists in more (costly) bookings for five-star hotel rooms in period 1 than in base period 0 for a given holiday destination. Even in the case of constant prices, a *unit value price index* would signal a price increase in period 1 simply related to the compositional changes in the hotels booked between both periods. Nevertheless, the *unit value price index* uses most of the transactions (see Figure 5) and is drift-free by construction in comparison with chain-linked price indices; therefore, it can serve as a simple benchmark method for the following (more sophisticated) price index methods.

## 4.2. Hedonic regression methods

Hedonics are a group of regression techniques, which describe the price of a given good or service as a function of several (observed) attributes, each having a marginal contribution to the overall price. In official statistics, hedonics are widely used in order to estimate a quality-adjusted price, for example in the context of residential house prices (see ILO et al. (2004); Triplett (2006); Eurostat (2013)). In the following, two different hedonic methods are tested with bookings of package holidays. The first method is *double imputation* (see Section 4.2.1), where prices are estimated for the base period as well as the comparison period. The second method is the *time dummy model* (see Section 4.2.2), where the index is directly derived from the coefficient of a time dummy variable in the regression.

([21]) Note that this implies a proportional relationship between the total price and both the number of days and the number of travellers. However, the price of a package holiday might be better reflected by a fixed-cost (travel-related) component and a non-proportional increase for additional travellers and/or days. This assumption is relaxed in the hedonic regression models in the next section.

### 4.2.1. DOUBLE IMPUTATION

Hedonic regression techniques can be used to estimate prices for products which are available in the base period 0 but are no longer available in the comparison period ***t***. To account for this fact, German official price statistics use the *double imputation* technique ([22]) for the house price index ([23]) and price indices of electronic products such as notebooks or smartphones, since the life cycle of innovative products is typically only a few months. Similarly, package holidays have a high churn, because they are rarely observed with exactly the same attributes in two successive periods. Some of the reasons for this are the numerous characteristics of package holidays as well as the seasonality of holiday destinations; for example, the number of bookings for Greece declines drastically during the winter season. Consequently, the *double imputation* is performed for package holidays on the basis of estimated prices for both the base period (year 2015) and a given comparison month ***t*** ([24]). Prices are estimated using the ordinary least squares method for the base year and for month ***t***. Consequently, the observations of month ***t*** are used to estimate prices for the base year (using the regression coefficients of the base year) and prices for month ***t*** (using the regression coefficients of month ***t***). In contrast to electronic products, the underlying regression model for package holidays is regarded as stable over a longer period of time, since the price-determining variables rarely change ([25]).

The Amadeus dataset contains several price-determining variables, as listed in Tables A.1 and A.2. In a first step, the variable selection of the regression model per holiday destination was done by analysing adjusted $R^2$ and its minimum and maximum range, indicating the explanatory content of the regression model. To avoid multicollinearity, the variance inflation factor (VIF) and significance of coefficients were checked. Moreover, the coefficients had to be stable and plausible over time, for example a coefficient of the four-star hotel dummy should be *ceteris paribus* smaller than the coefficient of the five-star hotel (see also Appendix A.3). Various combinations of variables were tested. For the variables ***travellerCount***, ***duration*** and ***bookTime***, three transformations were considered (continuous, log-transformation, or categorised), with the best option to use logarithmic values for all three variables. Moreover, in estimating a price properly, the *double imputation* method requires to capture the additional effect of public holidays — besides the typical holiday season — during a given travel month on the total price. Therefore, a dummy variable (***isHoliday***) is generated that equals 1 if Easter, Pentecost or Christmas falls during a given package holiday and 0 otherwise ([26]). Overall, a

([22]) Typically, the starting point for the concept of double imputation is an A-, B- and C-sample, where the B-sample contains all products that are present in both base period 0 and comparison period ***t***, and products of A- or C-sample are not present in either the base period (C-sample) or the comparison period (A-sample). However, the concept of the A-, B- and C-sample is not applicable for package holidays, since there is no B-sample available. See Linz et al. (2004) for further details on the double imputation technique applied by the Federal Statistical Office.

([23]) See Eurostat (2017), Section 6.1.2.

([24]) By contrast, for electronic products, January is chosen as a base period and the index is chain-linked annually. This allows an annual adjustment of the regression model to integrate new price-determining features. See Destatis (2009).

([25]) A change in the hedonic regression model for package holidays would only be necessary if, for example, the data provider changes the variables listed in Table A.1.

([26]) For example, the coefficient for ***isHoliday*** was 0.28 for the Canary Islands in December 2015, thus, the price of a package holiday is about 28 % higher for travelling at Christmas than for travelling before or after Christmas. Alternatively, one could also include public school holidays as an explanatory variable, although the date of these can vary considerably across the German Federal States.

model comprising variables ***travellerCount***, ***duration***, ***bookTime***, ***channel***, ***star*** and ***isHoliday*** gave the best results, with an average adjusted $R^2$ per holiday destination ranging between 0.704 and 0.785 (see Table A.4) ([27]). The final regression model comprising both online and offline bookings is subsequently defined as:

$$\begin{aligned}\ln\left(totalPrice_{i,t}\right) = {} & \beta_0 + \beta_1 \ln\left(travellerCount_{i,t}\right) + \beta_2 \ln\left(duration_{i,t}\right) + \\ & \beta_3 \ln\left(bookTime_{i,t}\right) + \beta_4 \mathrm{D}\left(channel_{i,t}\right) + \beta_5 \mathrm{D}\left(star_{one_{i,t}}\right) + \\ & \beta_6 \mathrm{D}\left(star_{two_{i,t}}\right) + \beta_7 \mathrm{D}\left(star_{three_{i,t}}\right) + \beta_8 \mathrm{D}\left(star_{five_{i,t}}\right) + \\ & \beta_9 \mathrm{D}\left(isHoliday_{i,t}\right) + \varepsilon_{i,t}\,,\end{aligned} \tag{3}$$

where Equation (3) is estimated for the base year 2015 and each comparison travel month ***t*** separately. Consequently, the Jevons formula is used for index calculation, in other words, the geometric mean of the estimated price relative of period *t* and base period ***0***, such that the index value for hedonic regression, $I_t^{DI}$, reads as follows:

$$I_t^{DI} = \left( \prod_{i=1}^{N} \frac{\hat{P}_{i,t}}{\hat{P}_{i,0}} \right)^{\frac{1}{N}}. \tag{4}$$

Note that ***mealType*** and ***roomCategory*** are also important price-determining variables, but are available for online bookings only. As a robustness exercise, a more detailed regression specification based on online transactions was estimated:

$$\begin{aligned}\ln\left(totalPrice_{i,t}\right) = {} & \beta_0 + \beta_1 \ln\left(travellerCount_{i,t}\right) + \\ & \beta_2 \ln\left(duration_{i,t}\right) + \beta_3 \ln\left(bookTime_{i,t}\right) + \beta_4 \mathrm{D}\left(star_{one_{i,t}}\right) + \\ & \beta_5 \mathrm{D}\left(star_{two_{i,t}}\right) + \beta_6 \mathrm{D}\left(star_{three_{i,t}}\right) + \beta_7 \mathrm{D}\left(star_{five_{i,t}}\right) + \\ & \beta_8 \mathrm{D}\left(seaView_{i,t}\right) + \beta_9 \mathrm{D}\left(highStandard_{i,t}\right) + \\ & \beta_{10} \mathrm{D}\left(lowStandard_{i,t}\right) + \beta_{11} \mathrm{D}\left(allInclusive_{i,t}\right) + \\ & \beta_{12} \mathrm{D}\left(breakfastOnly_{i,t}\right) + \beta_{13} \mathrm{D}\left(isHoliday_{i,t}\right) + \varepsilon_{i,t}\,,\end{aligned} \tag{5}$$

where additional dummy variables for the room and meal category were included. In the special case of Greece, due to a lack of bookings during the winter season, it is only possible to estimate a price index for the period May to October for each year ([28]).

([27]) The Federal Statistical Office also calculates other hedonic indices, which have an adjusted $R^2$ about 80 % (for complex products like servers) and nearly 100 % (for simple products like RAM modules). However, possible price-determining characteristics of package holidays such as hotel rating, hotel facilities or the exact location of a hotel are not available from the Amadeus dataset. Thus, an adjusted $R^2$ of about 0.75 for a complex product like package holidays seems to be acceptable.

([28]) To calculate a price index for the whole year for Greece, one possible solution would consist of a regression model with joint dummy variables for Greece and the Balearic Islands as Mediterranean euro area holiday destinations. However, the results were more plausible when using a single regression model for each holiday destination.

### 4.2.2. TIME DUMMY MODEL

The second hedonic method is the *time dummy model*, which also constitutes a regression approach. Contrary to the *double imputation* technique, no prices are estimated, but the price index is derived directly from the time dummy coefficient. For the *time dummy model*, the same regression model as in Equation (3) is taken, except for ***isHoliday***. The effect of public holidays has to be measured as a price change and is therefore already included in the time dummy variable ([29]). The *time dummy* regression model is given by:

$$\begin{aligned}\ln\left(totalPrice_{i,t}\right) &= \beta_0 + \beta_1 \ln\left(travellerCount_{i,t}\right) + \beta_2 \ln\left(duration_{i,t}\right) + \\ &\beta_3 \ln\left(bookTime_{i,t}\right) + \beta_4 \mathrm{D}\left(channel_{i,t}\right) + \beta_5 \mathrm{D}\left(star_{one_{i,t}}\right) + \\ &\beta_6 \mathrm{D}\left(star_{two_{i,t}}\right) + \beta_7 \mathrm{D}\left(star_{three_{i,t}}\right) + \beta_8 \mathrm{D}\left(star_{five_{i,t}}\right) + \gamma D_{i,t} + \varepsilon_{i,t}, \end{aligned} \tag{6}$$

where $D_{i,t}$ denotes the time dummy which equals 0 for the base period and 1 for the comparison travel month $t$ ([30]). The regression is estimated using all observations from the base period (January) and month $t$. The *time dummy model* index, $I_t^{TD}$, is directly derived from the exponential of the coefficient of the time dummy, $\gamma$, such that:

$$I_t^{TD} = e^{\hat{\gamma}}. \tag{7}$$

The final index series is chain-linked in January by applying the growth rate to the previous index value ([31]).

## 4.3. Stratification methods

An alternative to setting-up a regression model consists of dividing a sample into homogeneous strata and to consequently compute an average price within a given stratum. The following sections are dedicated to this *stratification approach*. As a first step, Section 4.3.1 deals with the definition of homogeneous strata or products in the context of package holidays by a quantitative approach. In a next step, Section 4.3.2 presents a traditional bilateral stratification approach based on a comparison of two periods, whereas Section 4.3.3 presents a multilateral approach, the *GEKS* method recently applied to supermarket scanner data, which compares several periods in computing a price index.

([29]) Including ***isHoliday*** in the *time dummy model* could also lead to multicollinearity, because both the time dummy variable and ***isHoliday*** measure a seasonal effect.
([30]) Here, we use only one time dummy and compare two periods, so the result is a bilateral index. It would also be possible to include more periods and extend the regression model by using more than one time dummy.
([31]) Hill (2011) suggests using a correction factor in the index calculation, because of a bias in the price index, which results from the fact that $E[e^{\hat{\gamma}}] \neq e^{\gamma}$. However, in the present application, the effect of the correction factor was quite small so the factor was not included in the final model.

### 4.3.1. PRODUCT DEFINITION BY A QUANTITATIVE APPROACH

In price statistics, a proper product definition is key. This is especially true for stratification methods as these methods group the underlying data according to their price-determining characteristics. Thereby, it is important to distinguish between items and products. More specifically, several items form one product ([32]). All items have certain characteristics of attribute variables and the question is which variables are important for product distinction and which ones can be neglected. Obviously, this problem is very much dependent on the product market and especially on the corresponding rate of churn.

Figure 6 illustrates the relationship between items and products in the context of package holidays. The right-hand column underlines the fact that the product definition at a lower level is not related to the index method since this calculation is performed at a higher aggregation level. In the second column, for illustrative purposes, some package holidays would form, for instance, the homogeneous product 'Turkish package holidays in three-star hotels, booked online within 30 to 90 days of departure'. This product again may form, along with several others, an ECOICOP subindex called 'Turkish package holidays'. Note that the Federal Statistical Office currently only publishes at a higher aggregation level (domestic and international package holidays). But if the sample covers a sufficient number of observations, it might also be feasible to publish subindices at a more detailed ECOICOP level such as by holiday destination to allow for a more detailed economic interpretation of the volatile prices of package holidays.

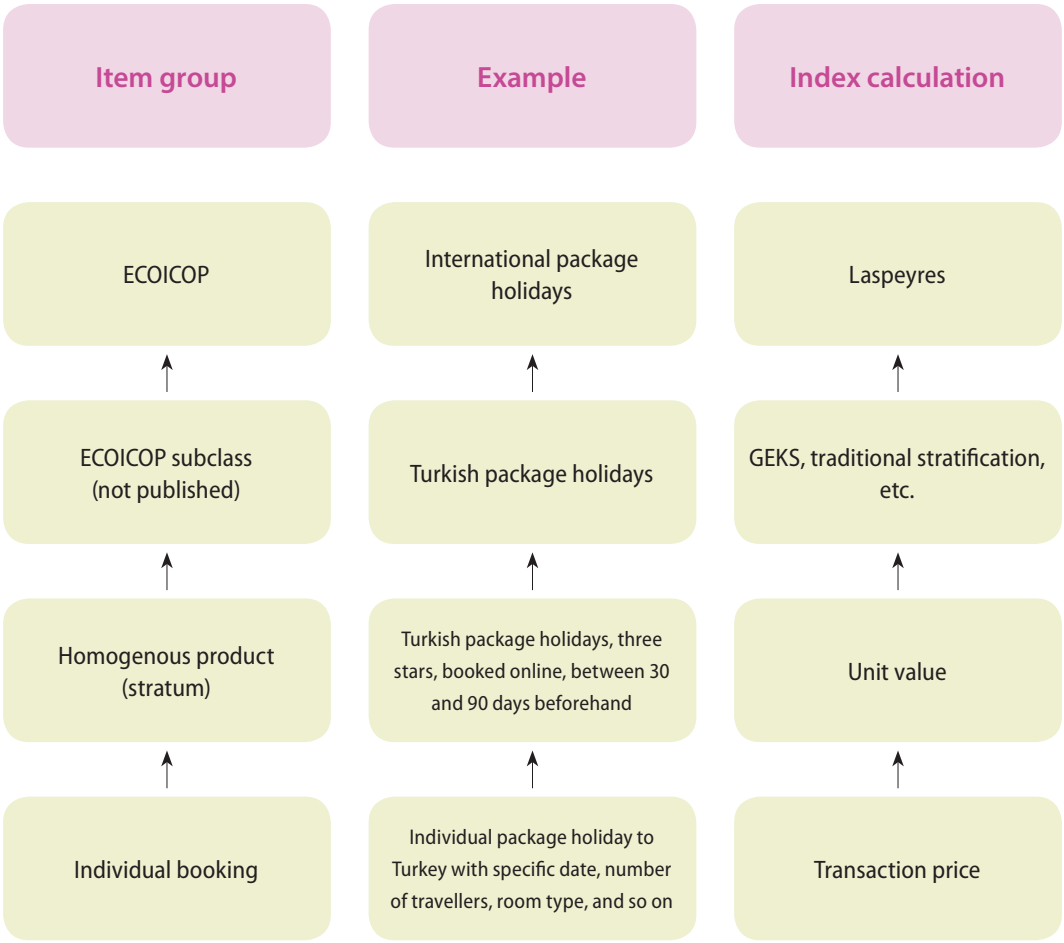
As a quantitative measure for the selection of price-determining variables for product definition, Chessa (2019) developed *Match Adjusted R Squared* (MARS). This measure weighs the two sides of product definition: product homogeneity and product continuity in comparison with a given base period. Thereby, product homogeneity among a specific product group is defined as the deviation of the average price, whilst assuming that homogeneous items do not vary much in price. Product continuity is defined as the share of products that are available in the base period as well as in the comparison period. Both measures are normalised to one. If, for example, a product definition is based only on the item level (in other words every single package holiday transaction), then product homogeneity equals one, but product continuity declines as new items appear on the market ([33]). Equally, if a product definition just aggregates all items to one product, the continuity is always one, but homogeneity would equal zero ([34]). Multiplying the values for product continuity and homogeneity yields the balance measure of MARS. This multiplication is similar to a classical loss function since product homogeneity increases as continuity decreases and vice versa.

([32]) A prominent application is in the field of clothing. While a single blue t-shirt of a certain brand with an individual Global Trade Item Number (GTIN) is an item, all blue t-shirts of any brand may form the product 'blue t-shirt', irrespective, for example, of the fabric or pattern. This product can be grouped again with t-shirts of other colours and other products to an ECOICOP subclass for 'men's shirts'.

([33]) This assumption is made implicitly for calculating the *double imputation* method (see Section 4.2.1) because no package holiday is grouped with another. Thereby, the lack of product continuity is handled by estimating the missing prices.

([34]) This assumption is made implicitly for the *unit value price index* (see Section 4.1) because no distinction between items is made.

**Figure 6:** Item hierarchy in the context of package holidays




*Source:* Own illustration following Chessa (2016)

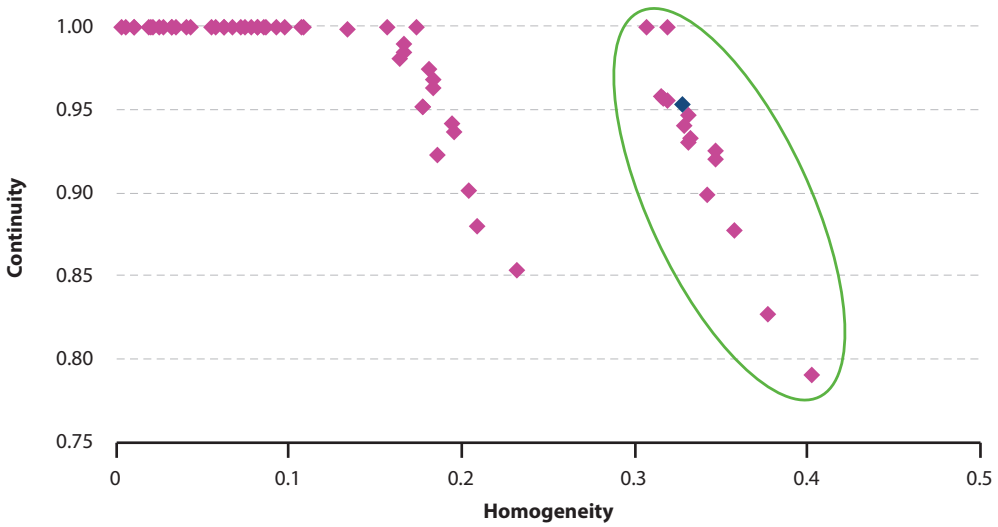
Applying this to package holidays, with ***n*** different product variables such as ***duration*** and accommodation category, there are $2^n$ different combinations forming a product definition at hand (not considering the number of attributes of a specific variable). By using ***PPD*** instead of ***totalPrice*** as the price variable, it is possible to omit two variables from the combinatorial problem (***duration*** and ***travellerCount***) ([35]). The variable ***bookTime*** was grouped in order to avoid a too detailed product definition ([36]). Moreover, since the shares of one- and two-star accommodations were relatively small in terms of the total revenue (for example less than 1 % and 2 % respectively in 2015), these bookings were removed beforehand. Likewise, the computation was only performed by using the 12 travel months for 2015, which also serves

([35]) Alternatively, transactions could be categorised by duration and the number of travellers. This would however largely increase the number of strata and therefore reduce product continuity.
([36]) As shown in Figure 4, the width of possible classes grows by increasing days before departure. Following this, the first group of ***bookTime_Class*** is from 15 to 30 days, the second from 31 to 90, the third class from 91 to 180, and the fourth class captures all bookings made more than 180 days in advance of departure.

as the base period for the following price indices. Using the results from hedonic regressions above as a starting point, six variables (***topArea***, ***star***, ***channel***, ***bookTime _Class***, ***depAirport*** and ***weekday*** of departure) were considered as variables for product definition ([37]). Thus, $2^6 = 64$ possible product definitions were tested.

Figure 7 depicts the average value of product continuity and homogeneity for the 64 tested product definitions in 2015. By concept, a combination in the upper right corner, where product continuity and homogeneity equal one, would be best (see the green circled set of points) ([38]). Moreover, a higher weight for product homogeneity seems to be more suitable for the heterogeneous product category of 'package holidays' ([39]). Based on the results in Table A.5 and the hedonic regression analysis before, a combination of variables is chosen which also exhibits a high number of items per product. In Figure 7, this combination is marked in dark blue, according to which a product in the context of package holidays is well defined by the variables ***topArea***, ***star***, ***channel*** and ***bookTime_Class***. Moreover, ***travellerCount*** and ***duration*** are included implicitly by using ***PPD*** rather than ***totalPrice*** as the price variable. Overall, these findings define the strata and the data filters that are used in the following two stratification methods.

**Figure 7: Continuity and homogeneity of several product definitions, 2015**




Note: following Chessa (2019).

*Source:* Bundesbank calculations on the basis of booking data from Amadeus Leisure IT GmbH; the product definition highlighted in dark blue was selected for subsequent analysis; for illustrative purposes, one outlier combination with homogeneity = 1 was excluded (in other words, every package holiday represented its own product class)

([37]) Note that seasonal variables like winter and summer season are not considered; they would create artificial breaks or discontinuities and therefore decrease the value of product continuity drastically. An alternative would be to stretch the base period to the entire previous year instead of just the previous month. However, this exercise is left for further research.

([38]) Note that it is not feasible to multiply the values for product continuity and homogeneity in Figure 7 in order to calculate MARS, since these represent averages from the 12 monthly values in 2015.

([39]) In the model from Chessa (2019), this can be thought as a loss function in an additive composition including a parameter λ for manual weighting.

### 4.3.2. TRADITIONAL STRATIFICATION

The *traditional stratification* approach tries to overcome the unit value bias of an average price by grouping transactions into several homogeneous classes before calculating the unit value. In terms of package holidays, transactions that have similar price-determining characteristics are sorted into the same class or stratum. In the following, for each holiday destination as given by ***topArea***, the strata are formed by ***star***, ***channel*** and ***bookTime_Class***, which is consistent with the set of variables approved by the results of the previous section and also the *hedonic regression*. The next step is to calculate in each stratum the average ***PPD*** in period ***t*** (see Equation (1)) and to normalise the resulting series to 2015 = 100 ([40]). In this manner, for each holiday destination, ***M***=24 strata are constructed resulting in 24 elementary price indices, $I^{TS}_{m,t}$.

The aggregation of those elementary price indices to an overall price index for the corresponding holiday destination can be affected by using either a weighted or unweighted mean. In some destinations, certain classes account for only a very small revenue share. For example, there tend to be less package holidays to three-star hotels in Turkey or five-star hotels on the Balearic and Canary Islands, respectively. Thus, an unweighted average price would be biased towards the under-represented classes. For this reason, the weighting is based on the total revenue shares of the individual class in 2015, as given by the transaction data. Finally, for each holiday destination the overall price index according to *traditional stratification*, $I^{TS}$, in period ***t*** is given by:

$$I^{TS}_t = \sum_{m=1}^{M} w_m I^{TS}_{m,t}, \tag{8}$$

where $w_m$ represents the 2015 revenue share of each stratum $m = 1,\ldots,M$.

In addition to the baseline version described above, two alternatives of *traditional stratification* are considered. First, bookings are grouped by ***iffCode***, which is the numeric identifier of the accommodation booked. Following this, the strata can be formed by ***iffCode***, ***channel*** and ***bookTime _Class*** ([41]). This selection of variables refines the baseline model above. Since the focus is now at the individual hotel level, the variable ***star*** can be neglected. For each of the six destinations, a large number of hotels are available, but concerning product continuity, it is reasonable to select the favoured ones. Therefore, for each holiday destination, only the top 25 hotels as measured by their revenue shares in 2015 are included in the calculation. Accordingly, the number of strata rises to ***M***=200, with 200 elementary price indices calculated and weighted together as described above. Second, by using only the online data, it is also possible to form the strata by using the variables ***star***, ***D(seaView)***, ***D(AllInclusive)***, and ***bookTime _Class*** to include price-determining information about meal and room categories ([42]). In this way, the resulting number of strata is ***M***=48, with 48 elementary price indices calculated and weighted in the same way as described above. For both alternatives, in case of missing bookings for a given period, the weights of the respective strata are set to zero and distributed proportionally across the remaining strata.

([40]) An additional stratification by ***duration*** and ***travellerCount*** would also be possible (for example one strata for 7-day package holidays and another one for 14-day package holidays). As a result, ***totalPrice*** could be used as the relevant price variable instead of ***PPD***. However, this would strongly reduce the number of observations per stratum.

([41]) It is also reasonable to stratify by ***iffCode***, ***bookTime_Class*** and ***depAirport*** (using the three largest German airports, for instance). However, the results were very similar to the stratification by ***iffCode***, ***channel*** and ***bookTime_Class***.

([42]) To cover meal type, only the variable ***D(AllInclusive)*** is included because some meal categories such as "breakfast only" would have none to very low observations for some holiday destinations. Regarding room category, it is reasonable to use only the indicator variable for sea view to guarantee a sufficiently high number of observations. For the same reason, the variable ***star*** is included instead of ***iffCode***.

### 4.3.3 GEKS

The origin of the following method goes back to Gini, Eltetö, Köves and Szulc (*GEKS*) and was adopted by Ivancic, Diewert and Fox (2011) to apply to the growing field of scanner data in price statistics ([43]). As in the previous approach, the price variable is ***PPD*** and the sample is stratified to calculate a unit value per stratum. The difference between *GEKS* and the *traditional stratification* approach lies in the index aggregation; instead of using the fixed weights from the year 2015, the monthly revenue shares of each stratum were used. Moreover, *GEKS* is a multilateral method, which compares more than two time periods to each other in computing a price index. The main advantage from multilateral methods is that these are transitive and therefore generally free from chain drift ([44]).

In particular, in the current month $T$, *GEKS* compares all months $t = 1,\ldots,T$ with the base month 0 using a geometric mean of a set of index ratios comprising the Fisher index of month 0 divided by the Fisher index of month $T$ whereas the base period iterates from 0 to $T$ ([45]). Given any period $t = 0,\ldots,T$, the *GEKS* index between base period 0 and comparison period $t$, $I_{0,t}^{GEKS}$, is defined by:

$$I_{0,t}^{GEKS} = \prod_{z=0}^{T} \left( \frac{P_{0,z}^{Fish}}{P_{t,z}^{Fish}} \right)^{\frac{1}{T+1}}, \tag{9}$$

where $I_{t,z}^{Fish}$ represents the Fisher index between period $t$ and $z$, whereas $T$ stands for the current month. The Fisher index is given by:

$$I_{t,z}^{Fish} = \sqrt{I_{t,z}^{L} \cdot I_{t,z}^{Pa}} = \sqrt{\frac{\sum_{i=1}^{N_{t,z}} p_z^i q_t^i}{\sum_{i=1}^{N_{t,z}} p_t^i q_t^i} \cdot \frac{\sum_{i=1}^{N_{t,z}} p_z^i q_z^i}{\sum_{i=1}^{N_{t,z}} p_t^i q_z^i}}, \tag{10}$$

with $I_{t,z}^{L}$ as the Laspeyres index and $I_{t,z}^{Pa}$ as the Paasche index between period $t$ and $z$. Furthermore, $p_t^i$ and $q_t^i$ denote the price and the quantity of product $i$ sold in month $t$. Lastly, $N_{t,z}$ stands for the total number of products that are sold in month $t$ as well as in month $z$. As reflected in Equation (9), multilateral indices inherit ongoing revisions; in the next period $T+1$, the value of $I_{0,t}^{GEKS}, (t = 0,\ldots,T)$ might be different to its value in period $T$ since the product is expanded by one factor. To avoid revisions of already published price indices, Ivancic et al. (2011) propose a chain-link. This is done by recalculating the indices for all other months with the help of the new month and applying the growth rate of the new month to the previously published index values (so-called movement splice). Additionally, the authors propose a rolling window in order to give more recent index values a higher weight in the current index calculation. Hence, $T$ reflects also as the size of the rolling window. In the present application, the length of the rolling window was set to 13 months ([46]). Note that no dumping-filter was applied, because data cleansing was done beforehand (see Section 3) ([47]).

([43]) Introduced already in the mid-1960s, this index concept is also used to measure purchasing power parities (see OECD and Eurostat (2012) for an overview).
([44]) Note that also hedonic regression methods can in principle be constructed in a multilateral way, which is, however, not the case in this paper.
([45]) Note that instead of a Fisher index, a Törnqvist index could also be applied.
([46]) See Van Loon and Roels (2018) for an overview of different chain-linking methods. Besides the movement splice, the fixed base moving window proposed by Lamboray (2017) was tested. The results were very similar. Following de Haan and Krsinich (2018), a window length of 13 months is the smallest that can deal with seasonal products. In the present case, the window initially starts in January 2014 and ends in January 2015 (for Greece, from May to October 2014 and May 2015).
([47]) In German price statistics, the *GEKS* method was applied by Bieg (2019) to supermarket scanner data.

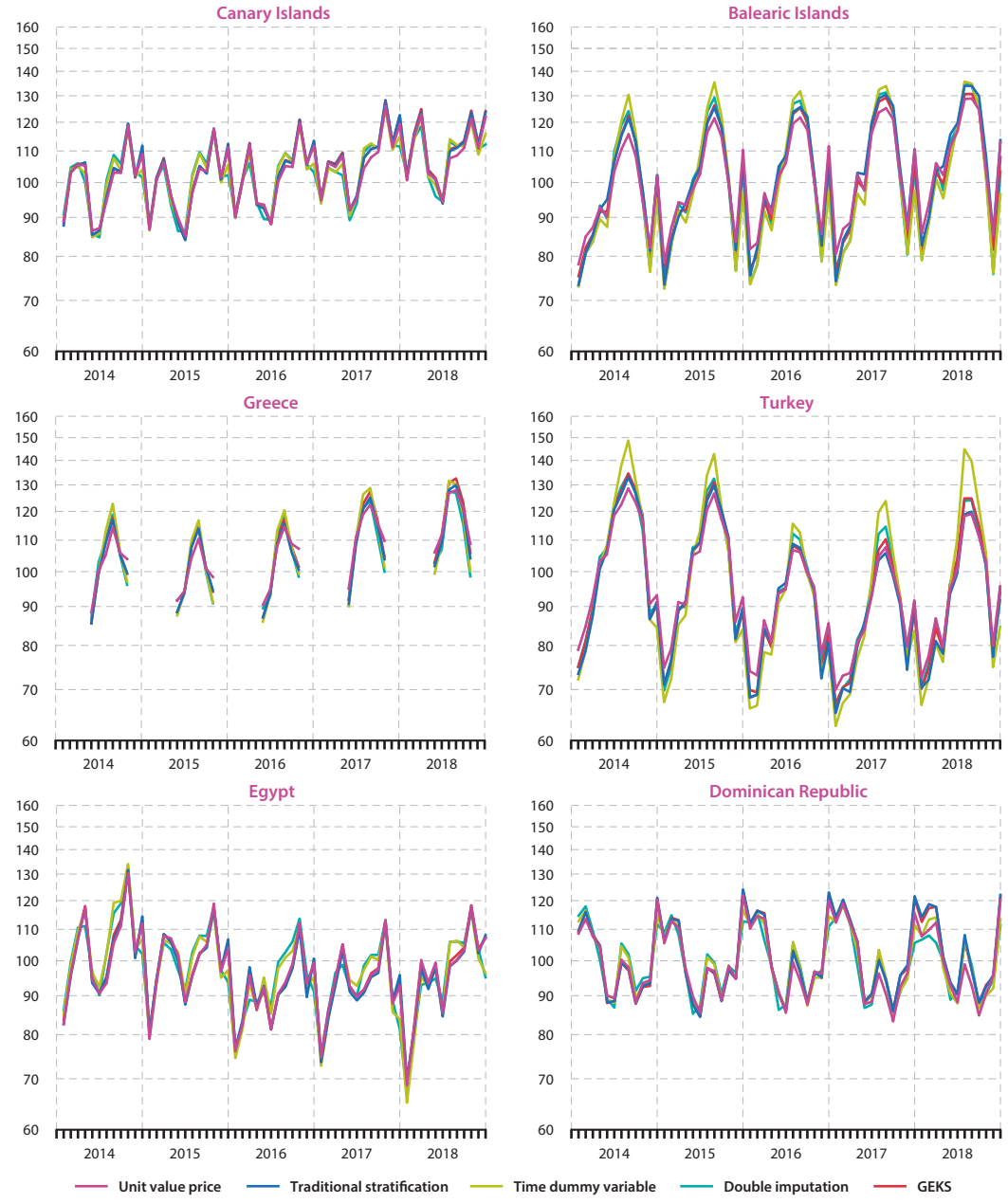
# 5. Comparison of results

In the following, price indices based on the five different methods (*unit value price index*, *double imputation*, *time dummy model*, *traditional stratification* and *GEKS*) are evaluated concerning their overall seasonal pattern, volatility and robustness with respect to different data filters. Ideally, all price indices follow a similar pattern for a given holiday destination, so that to a large extent the selection of the method does not influence the overall movement of the series. In this case, the decision on the preferred method could in principle be based on the volatility of the annual rates of change. Moreover, the resulting transaction-based price indices are compared with the official price index, which uses offer prices. For this purpose, for each method, the underlying dataset excludes last minute bookings (***bookTime*** ≤ 14) as well as non-German departure airports (***D(GermanAirport)*** = 0), which is consistent with the current official practice as described in Section 2.

Figure 8 shows transaction-based price indices according to the different methods for the six major holiday destinations (the Balearic and Canary Islands, Turkey, Greece, Egypt, and the Dominican Republic). Overall, the resulting price indices for package holidays in a given destination have the same seasonal pattern, with typically higher prices during summer months in Germany and lower prices during winter months. However, there are some differences across methods within specific destinations. For instance, at the end of each calendar year, the price trend for the Canary Islands based on *double imputation* differs from the price trends for the other methods. For Egypt, both methods of *hedonic regression* differ at the end of the year. For Turkey, the *time dummy model* exhibits a higher volatility in comparison with the other methods. For the Dominican Republic, the fourth quarter of 2017 and the first quarter of 2018 show differences between almost all methods. Note that although the concept of the *GEKS* as a multilateral index is very different from the bilateral indices, it provides similar results.

To have a closer look at the differences in dynamics between methods, the next step is to analyse the annual rates of change, in other words the percentage change between a given month and the same month of the previous year. For this purpose, descriptive statistics are calculated for each method and holiday destination. The *unit value* approach is generally less volatile; however, it is also considered to exhibit the lowest degree of product homogeneity over time (see Figure 5). The arithmetic mean (MEAN) indicates whether the price indices have the same trend over time, whereas the standard deviation (SD) as well as the minimum (MIN) and the maximum (MAX) indicate the volatility of the annual rates of change. In Table 2, the (absolute) lowest SD, MIN and MAX of a given holiday destination are highlighted in green. At a first glance, *traditional stratification* and *double imputation* perform well in terms of these descriptive statistics. The latter exhibits the lowest volatility as indicated by the standard deviation. However, it also appears that the performance of each method seems to depend on the holiday destination under consideration. Whereas for the Canary Islands and Egypt, *double imputation* performs best, in the Balearic Islands and Greece, *traditional stratification* seems to perform well. Note that the largest variation across methods is found for the Dominican Republic, where — in contrast to the other holiday destinations — the sign of the average rate of change (MEAN) differs between methods.

**Figure 8: Comparison of different methods of price measurement**
(2015 = 100, log scale)




*Source:* Bundesbank calculations on the basis of booking data from Amadeus Leisure IT GmbH

**Table 2: Descriptive measures of different index methods by holiday destination**
(annual rate of change, %)

| | | Unit value price | Hedonic regressions | | Stratification | |
|---|---|---|---|---|---|---|
| | | | Double imputation | Time dummy variable | Traditional stratification | GEKS |
| **Canary Islands** | Mean | 2.4 | 2.1 | 2.3 | 2.6 | 2.8 |
| | Standard deviation | 4.7 | 3.8 | 4.8 | 5.0 | 5.0 |
| | Min | -8.7 | -6.7 | -8.0 | -9.6 | -9.3 |
| | Max | 17.1 | 14.7 | 18.4 | 17.5 | 18.1 |
| **Balearic Islands** | Mean | 3.2 | 2.7 | 2.3 | 3.3 | 2.5 |
| | Standard deviation | 4.7 | 6.4 | 5.3 | 4.3 | 4.6 |
| | Min | -8.4 | -12.8 | -8.5 | -7.8 | -8.4 |
| | Max | 19.5 | 25.8 | 20.5 | 18.1 | 19.0 |
| **Greece** | Mean | 3.4 | 2.4 | 2.6 | 3.2 | 3.6 |
| | Standard deviation | 5.5 | 5.8 | 6.0 | 5.3 | 5.7 |
| | Min | -5.9 | -9.1 | -6.1 | -5.5 | -5.7 |
| | Max | 16.9 | 18.2 | 18.6 | 16.0 | 18.2 |
| **Turkey** | Mean | -2.1 | -2.1 | -1.8 | -2.2 | -1.9 |
| | Standard deviation | 8.2 | 8.3 | 9.9 | 8.8 | 9.1 |
| | Min | -16.3 | -16.8 | -21.0 | -17.4 | -18.5 |
| | Max | 17.9 | 19.6 | 21.0 | 16.6 | 18.2 |
| **Egypt** | Mean | -1.5 | -2.3 | -2.6 | -1.7 | -1.6 |
| | Standard deviation | 7.9 | 7.0 | 8.0 | 7.3 | 7.9 |
| | Min | -19.4 | -15.9 | -18.1 | -17.6 | -18.5 |
| | Max | 21.8 | 16.9 | 20.8 | 18.2 | 20.9 |
| **Dominican Republic** | Mean | -0.3 | -0.5 | -0.3 | 0.7 | 0.6 |
| | Standard deviation | 3.4 | 3.3 | 3.0 | 3.4 | 3.2 |
| | Min | -7.7 | -8.8 | -7.1 | -7.9 | -8.1 |
| | Max | 8.4 | 5.5 | 5.8 | 7.9 | 7.8 |

Note: based on the period from January 2014 to December 2018.

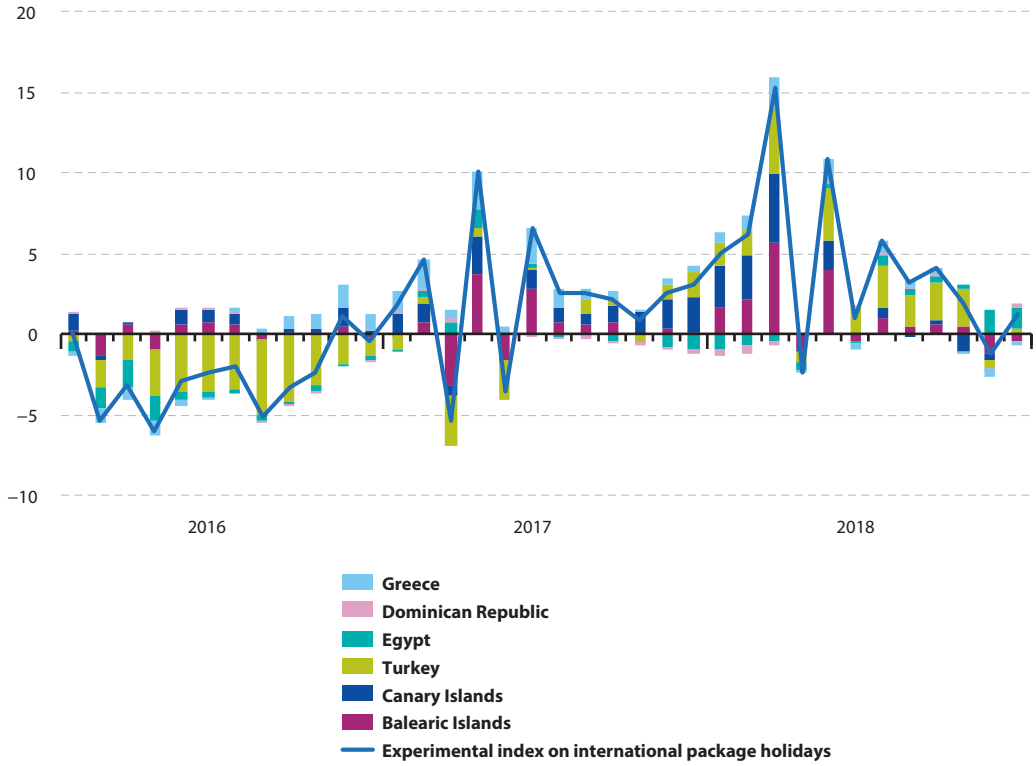
In addition, several robustness tests related to the data itself as well as to the model specifications were performed. Results for different data filters are shown in the Appendix for *double imputation* and *traditional stratification* (Figures A.2 and A.3, respectively) ([48]). Using all of the transaction data including last minute bookings (those made 14 days or less before departure) as well as including non-German departure airports did not affect the index values in a noticeable way. Moreover, excluding bookings with an accompanying child (aged less than 16 years), which might comprise a tour operator-specific discount on the package holiday, did not impact the resulting series. By using online bookings only, a more detailed regression Equation (5) was estimated for the *double imputation* method (see Section 4.2.1). Evidently, using the additional information on the meal type (for example 'all inclusive') or the room category does not seem to change the resulting hedonic price index. Finally, two alternatives of the *traditional stratification* approach were tested: a more detailed stratification for online bookings by including also the information on the meal type and room category as well as a stratification at the individual hotel level (see Section 4.3.2). Whereas the resulting annual rates of change of the first alternative closely resemble the rates of the baseline version of *traditional stratification*, stratification at the individual hotel level differs quite greatly in some

([48]) For a detailed description of the robustness exercise concerning different datasets, see Table A.6.

periods and is also more volatile (see Figure A.4 in the Appendix). For the Balearic Islands, the resulting rates of change deviate notably during winter months. This is due to the sharp drop in observations, noting that hotels booked during the winter season are generally different to those booked during the summer season ([49]). Similarly, the annual rates of change for Turkey differ widely during the winter season of 2017/2018. Overall, the number of observations for these alternative specifications decreases considerably; in comparison with the baseline versions, only one quarter to one third of the data is used (see Table A.7 in the Appendix). Therefore, in the remaining analysis, only the baseline versions of the *double imputation* and the *traditional stratification* method are considered.

The compilation of destination-based price indicators allows for a detailed economic interpretation of the overall price trend for international package holidays. In this sense, Figure 9 plots an experimental price index based on the baseline version of the *double imputation* method, by aggregating the six destination-based price indicators using their average revenue shares from 2015-2016. It becomes clear that the negative price trend in 2016

**Figure 9: Experimental index for international package holidays and contributions from holiday destinations**
(percentage points for the contribution of holiday destinations, % change compared with same month of the previous year for the experimental index)




*Source:* Bundesbank calculations on the basis of booking data from Amadeus Leisure IT GmbH; the experimental index is based on the double imputation method and consists of six holiday destinations, which are weighted together by their respective revenue share (2015-2016 average)

([49]) Most of the top 25 hotels have no observations in the winter months (November to February). The remaining top 25 hotels have only a small proportion of their observations during this period. This leads to 'unusual' prices and accordingly to more volatile rates of change.

as well as the recent peak in the summer of 2018 in international package holiday prices was primarily driven by developments in Turkey. During the beginning of the sample, the latter experienced a decline in bookings as a response to several terroristic attacks and increasing political uncertainty, with bookings recovering in the summer season of 2018. Obviously, this was accompanied by a similar movement in prices for package holidays in Turkey. Due to the resulting shift in German travellers' preferences, the Balearic and Canary Islands and, to a lesser extent, Greece, could at the same time increase their prices for package holidays during 2017 and 2018 ([50]).

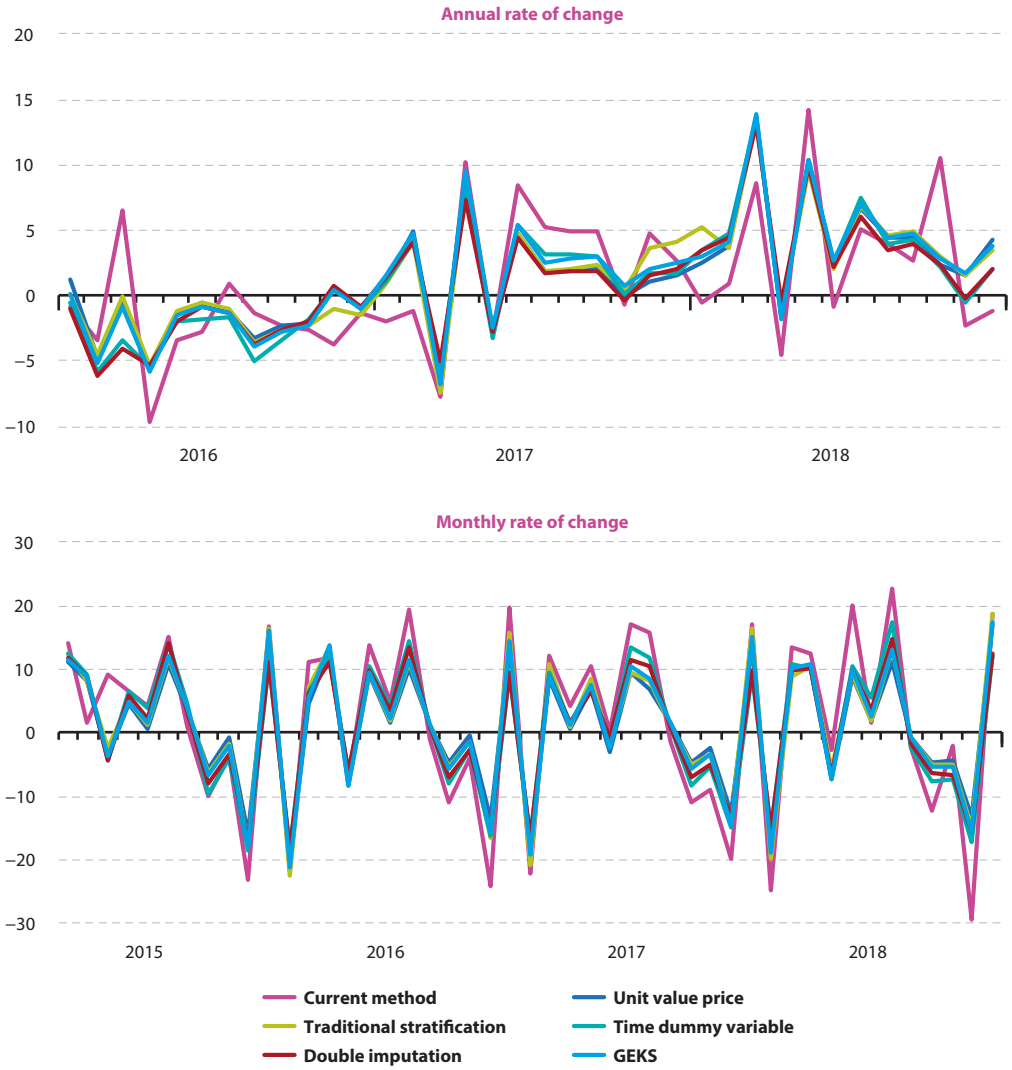
Finally, the transaction-based indices are contrasted with the official price index for international package holidays (ECOICOP 09.6.0.2), which is based on offer prices and is currently only reported at the aggregate level. For an approximate comparison, the transaction-based indices for the six holiday destinations are used to calculate an overall index for package holidays abroad according to the calculation procedure of the official index ([51]). As described in Section 2, the official price index consists of these six holiday destinations for international flight package holidays, but also includes city trips and cruises. The latter two are not calculated with Amadeus transaction data; instead, the official (confidential) subindices are used ([52]). Similarly, transaction-based indices for Greece and cruises during the winter season are imputed by using all available subindices (all-seasonal estimation). For the Dominican Republic, the official subindex imputes the summer months whereas the transaction-based indices for this holiday destination are also based on actual bookings during the summer season ([53]). For all five transaction-based methods under consideration, a corresponding index for international package holidays is calculated by summing up the eight subindices using the official weighting scheme.

Figure 10 depicts the annual rates of change for all five transaction-based indices together with the current official index. Note that a comparison of the latter can be only made from January 2016 onwards, since a new computation method was introduced (with data back to January 2015). Concerning the annual rates of change as shown in the upper part of Figure 10, there are only four periods (out of a total of 36), when the algebraic sign of the respective rate of change diverges across the five transaction-based methods. In contrast, the official method deviates in 11 out of the 36 periods from the sign for the rate of change indicated by the majority of transaction methods. Concerning month-on-month rates of change from February 2015 onwards, the five methods do not differ in any of the 47 periods in terms of their signs for the rate of change, reflecting the dominance of the seasonal pattern in the series. The official method deviates only in four out of the 47 periods. Finally, descriptive statistics for the annual rates of change are shown in Table 3. Evidently, all methods have a smaller standard deviation when compared with the official method. Concerning the different indices, *double imputation* has the lowest standard deviation; however, the differences between methods fall within a rather small range.

([50]) See also Section 3 on revenue shares per holiday destination over time. Note that, in calculating the contributions to growth, the weight of a given holiday destination was held constant (average 2015-2016 revenue share).
([51]) Note that the official weighting scheme at this detailed level is not published.
([52]) Concerning cruises, in the transaction data there is only information on the destination airport, but not on the room category (for example, inside or outside cabin), which is obviously an important price determinant when booking a cruise. City trips might be calculated with the Amadeus data, but this is left for further research.
([53]) This does not only affect price movements for the Dominican Republic but also indices for Greece and cruises, because out-of-season months are imputed using the all-seasonal estimation.

**Figure 10: Comparison of transaction-based pseudo indices with the current subindex for international package holidays (ECOICOP 09.6.0.2)**

(%)




*Source:* Bundesbank calculations on the basis of booking data from Amadeus Leisure IT GmbH; the experimental index is based on the double imputation method and consists of six holiday destinations, which are weighted together by their respective revenue share (2015-2016 average)

**Table 3: Comparison of transaction-based methods with current national practice**
(% change compared with the same month of the previous year)

| | HICP international package holidays (09.6.0.2) | Unit value price | Double imputation | Time dummy variable | Traditional stratification | GEKS |
|---|---|---|---|---|---|---|
| Mean | 4.3 | 3.1 | 3.2 | 3.5 | 3.3 | 3.4 |
| Standard deviation | 5.3 | 4.3 | 4.1 | 4.5 | 4.3 | 4.4 |
| Min | -9.7 | -6.4 | -6.1 | -5.8 | -7.5 | -6.7 |
| Max | 14.3 | 13.7 | 13.1 | 13.6 | 13.4 | 14.0 |
| First quartile | -2.5 | -1.4 | -1.8 | -1.8 | -1.2 | -1.5 |
| Third quartile | 4.9 | 4.2 | 3.5 | 3.8 | 4.2 | 4.0 |

Note: based on the annual rates of change from January 2016 to December 2018, since the method of the official price index changed in 2015. For each transaction-based method, the elementary indices for six holiday destinations were aggregated together with the official (confidential) elementary indices for 'city trips' and 'cruises' by using the official weighting scheme.

All in all, the transaction-based methods presented above generate similar price indices, which do not vary a lot over time. This is in contrast to the current method that is based on offer prices, where differences compared with the transaction-based methods become apparent during certain periods (see Figure 10). The reasons for these differences are hard to judge. One reason might be the different underlying principles of price comparison. The current official method is based on a pure price comparison of identical price offers over time by tracking the same booking code in each month, in other words quality changes should not influence price developments. Methods that are based on transaction data also try to compare like with like but define identical products for package holidays in a broader way ([54]). Thus, transaction-based methods might not eliminate heterogeneity in bookings to a sufficient degree and might therefore suffer from model uncertainty caused by structural shifts and substitution effects. Moreover, price collection that is based on offer prices might be prone to sampling uncertainty as in the case of every statistic that is based on samples. In that sense, the transaction-based indices cover a more universal dataset by using approximately 50 to 100 times more observations per year than the current official practice (see Table A.7).

# 6. Summary

This paper has shown that, by means of transaction data, it is possible to calculate efficiently several experimental price indices that can be disaggregated by holiday destination, therefore allowing the interpretation of movements in the overall price index for international package holidays. All five methods under consideration follow a similar pattern, from which the official price index based on offer prices deviates at some points in time.

Concerning the difference between transaction-based and offer-based methods, there remain some open questions. In comparison with offer prices, it is not clear to what extent the given transaction-based methods perform sufficiently well in terms of varying sample and

([54]) Note that there is currently an on-going discussion in price statistics about the appropriate definition of 'homogeneous products' being a challenge when using new digital data sources. See, for instance, Zhang et al. (2019) as well as Nilsson and Ståhl (2019).

quality adjustment, notably regarding incomplete information such as the exact room type. Whereas transaction-based methods might suffer from 'model uncertainty', there is always a potential 'sample uncertainty' when using offer prices. Moreover, it is not sure whether the sampled offer prices represent a transaction. A quantification of both effects has to be based on a comparison between transaction prices and offer prices at the level of individual bookings, which is beyond the scope of this paper. Note that it would also be fruitful to extend research on measuring prices to cruises as these are thought to be an important driver of price developments in the German package holiday market; this would require more detailed information, for example on the cabin category booked. Moreover, transaction data from other global distribution systems or even from tour operators themselves could make the analysis more robust.

Concerning an implementation of the current transaction-based methods in statistical production and the publication of destination-based price indicators, several important issues have to be noted. If one states that a pure price comparison can only be achieved via Laspeyres-like methods, then some of the methods presented are not fully in line with the current HICP regulation. The *GEKS* method applied in this paper relies heavily on a Fisher index that uses changing weights due to the underlying Paasche index. Nevertheless, Eurostat is currently working on adapting the current legal framework to allow for other price index formulae beyond Laspeyres. Finally, concerning a more detailed breakdown of the German HICP for package holidays, note that this aim might also be accomplished with offer data. The Federal Statistical Office is currently extending their price collection to a larger number of price representatives per destination and to a larger number of travel days per month by means of an automated interface to the Amadeus booking system. Hence, a future disaggregation by holiday destination could also be developed based on offer data. Nonetheless, in the case of offer data, collected prices still need to be aggregated using external weight information, for example survey or also transaction data concerning the time of booking. In this sense, transaction data, which already contain weight information on a very detailed level, might be more convenient.

# Acknowledgements

The authors would like to thank two anonymous reviewers for their valuable comments and suggestions. The authors would also like to thank Timm Behrmann, Dorothee Blang, Edgar Brandt, Antonio Chessa, John Johansson, Thomas Knetsch, Michael Kuhn, Susanne Lorenz and Jens Mehrhoff as well as seminar participants at the Deutsche Bundesbank, at the 23rd Workshop on Price Measurement held by the Statistical Office Berlin-Brandenburg and at the 16th Meeting of the Ottawa Group on Price Indices in Rio de Janeiro for their helpful comments and discussions.

# References


Bieg, M. (2019), 'Nutzung von Scannerdaten in der Preisstatistik – Eine Untersuchung anhand von Marktforschungsdaten' (in German only) *WISTA — Wirtschaft und Statistik*, No. 2/2019, pp. 25-38.

Chessa, A.G. (2016), 'Processing Scanner Data in the Dutch CPI: A New Methodology and First Experiences', *Eurostat Review on National Accounts and Macroeconomic Indicators (EURONA)*, 1/2016, pp. 49-69.

Chessa, A.G. (2019), 'MARS: A Method for Defining Products and Linking Barcodes of Item Relaunches', 16th Ottawa Group meeting, Rio de Janeiro, Brazil, 8-10 May 2019.

Destatis (2009): *Handbook on the Application of Quality Adjustment Methods in the Harmonised Index of Consumer Prices*, Statistics and Science, Volume 13, Federal Statistical Office of Germany.

Deutsche Bundesbank (2017), 'The Volatility of the Traditional Core Inflation Rate in Germany', Monthly Report, November 2017, pp. 49-51.

Deutsche Bundesbank (2019), 'The Revision of the Sub-Index for Package Holidays and its Impact on the HICP and Core Inflation', Monthly Report, March 2019, pp. 8-9.

European Commission (2009), Commission Regulation (EC) No 330/2009 of 22 April 2009 laying down detailed rules for the implementation of Council Regulation (EC) No 2494/95 as regards minimum standards for the treatment of seasonal products in the Harmonised Indices of Consumer Prices (HICP).

Eurostat (2013), *Handbook on Residential Property Price Indices (RPPIs)*, Publications Office of the European Union, Luxembourg.

Eurostat (2017), *Technical Manual on Owner-Occupied Housing and House Price Indices*, Luxembourg.

Eurostat (2018), *HICP Methodological Manual*, Publications Office of the European Union, Luxembourg.

Eurostat (2019), 'Improved Calculation of HICP Special Aggregates and German Package Holidays Methodological Change', press release, 22 February 2019 (updated, 1 March 2019).

de Haan, J. and F. Krsinich (2018), 'Time Dummy Hedonic and Quality-Adjusted Unit Value Indexes: Do They Really Differ?', *Review of Income and Wealth*, Vol 64, Issue 4), pp. 757-776.

Hill, R. (2011), 'Hedonic Price Indexes for Housing', OECD Statistics Working Paper No. 2011/01.

ILO, IMF, OECD, UNECE, Eurostat, and World Bank (2004), *Consumer Price Index Manual: Theory and Practice*, ed. Hill, P., International Labour Office, Geneva.

IMF (2009), *Export and Import Price Index Manual: Theory and Practice*, International Monetary Fund, Washington, D.C.

Ivancic, L., W. E. Diewert and K. J. Fox (2011), 'Scanner Data, Time Aggregation and the Construction of Price Indexes', *Journal of Econometrics*, Vol 161, Issue 1, pp. 24-35.

Johansson, J. and C. Tongur (2019), 'Package Holidays: Transaction Data in the Swedish CPI from 2019', technical note, *mimeo*.

Lamboray, C. (2017), 'The Geary Khamis Index and the Lehr Index: How Much Do they Differ?', 15th Ottawa Group meeting, Eltville am Rhein, Germany, 10-12 May 2017.

Linz, S., T. Behrmann and U. Becker (2004), 'Hedonische Preismessung bei EDV-Investitionsgütern' (in German only), *WISTA — Wirtschaft und Statistik*, No. 6/2004, pp. 682-689.

Nagengast, A., D. Bursian, J.-O. Menz (2019), 'Dynamic Pricing and Exchange Rate Pass-Through: Evidence from Transaction-Level Data', Deutsche Bundesbank discussion paper, forthcoming.

Nilsson, P. and O. Ståhl (2019), 'Towards a Roadmap for Efficient Use of Electronic Transaction Data in the Swedish CPI', 16th Ottawa Group meeting, Rio de Janeiro, Brazil, 8-10 May 2019.

OECD and Eurostat (2012), *Eurostat-OECD Methodological Manual on Purchasing Power Parities*, Publications Office of the European Union, Luxembourg.

Triplett, J. (2006), *Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes: Special Application to Information Technology Products*, OECD Publishing, Paris.

Van Loon, K. and D. Roels (2018), 'Integrating Big Data in the Belgian CPI', Meeting of the Group of Experts on Consumer Price Indices, Geneva, Switzerland, 7-9 May 2018.

Zhang, L.-C., I. Johansen and R. Nygaard (2019), 'Evaluating Unit-Value Price Indices in a Dynamic Item Universe', 16th Ottawa Group meeting, Rio de Janeiro, Brazil, 8-10 May 2019.

# Appendix

## A.1. Overview of variables

**Table A.1:** Description of variables in Amadeus dataset

| Variable | Description | Online | Offline | Type |
|---|---|---|---|---|
| **Information on accommodation** | | | | |
| iffCode | Numeric identifier of the accommodation booked | Y | Y | Numeric |
| accomCategory | Classification of the standard of the accommodation (star rating) | Y | Y | Numeric |
| accomName | Name of accommodation (for example, 'Sea View Hotel') | Y | Y | Alphanumeric |
| isCruise | Accommodation represents a cruise ('Y' or 'N') | Y | Y | Categorical |
| **Information on holiday destination** | | | | |
| accomLocation | Location (lowest level of geography) of the accommodation (for example, Playa de Palma) | Y | Y | Alphanumeric |
| accomProvince | Area of the accommodation (for example, Balearic Islands) | Y | Y | Alphanumeric |
| accomCountry | Country of the accommodation area (for example, Spain) | Y | Y | Alphanumeric |
| **Information on flight** | | | | |
| travelDate | Date on which travel is booked to start | Y | Y | Date |
| depAirport | 3-letter IATA code of departure airport | Y | Y | Alphanumeric |
| destAirport | 3-letter IATA code of destination airport | Y | Y | Alphanumeric |
| **Information on booking process** | | | | |
| tourOperatorId | Numeric identifier of tour operator | Y | Y | Numeric |
| channel | Source of the booking ('online' or 'offline') | Y | Y | Categorical |
| status | Status of the booking ('booked' or 'cancelled') | Y | Y | Categorical |
| transactionDate | Date on which the booking is made | Y | Y | Date |
| postcode_travelAgency | Post code of traditional high street travel agency | N | Y | Numeric |
| **Information on travellers** | | | | |
| travellerCount | Number of travellers on the booking | Y | Y | Numeric |
| childrenCount | Number of children and infants on the booking | N | Y | Numeric |
| travellerAges | List of ages of each of the travellers | Y | N | Alphanumeric |
| **Information on transaction price** | | | | |
| totalPrice | The selling price of the booking expressed in EUR | Y | Y | Numeric |
| duration | Length of the holiday expressed as a number of days | Y | Y | Numeric |
| mealType | A classification of the level of service provided at the accommodation (for example, 'all inclusive') | Y | N | Alphanumeric |
| roomCategory | Description of the accommodation booked (for example, 'with sea view') | Y | N | Alphanumeric |
| hasTravelInsurance | Total price includes travel insurance ('Y' or 'N') | Y | N | Categorical |
| hasHireCar | Total price includes car hire ('Y' or 'N') | Y | N | Categorical |

**Table A.2:** Description of newly defined variables

| Variable | Description | Type |
|---|---|---|
| travelMonth | Month of travelDate | Numeric |
| bookingMonth | Month of transactionDate | Numeric |
| bookTime | Difference between travelDate and transactionDate in number of days | Numeric |
| bookTime_Class | bookTime divided into four classes (up to 30 days, between 31 and 90 days, between 91 and 180 days, higher than 180 days) | Numeric |
| PPD | Price per person per day | Numeric |
| children | Number of children (offline) and travellers aged less than 16 years (online) | Numeric |
| star | accomCategory divided into five classes (one to five stars) | Numeric |
| D(star_one) to D(star_five) | Dummy variables for a given star category (1 or 0) | Categorical |
| D(online) | Online booking only (1 or 0) | Categorical |
| D(GermanAirport) | destAirport is located in Germany (1 or 0) | Categorical |
| topArea | Balearic Islands, Canary Islands, Turkey, Greece, Egypt or Dominican Republic | Alphanumeric |
| D(doubleRoom) | Indicator variable (see Table A.3) | Categorical |
| D(seaView) | Indicator variable (see Table A.3) | Categorical |
| D(highStandard) | Indicator variable (see Table A.3) | Categorical |
| D(lowStandard) | Indicator variable (see Table A.3) | Categorical |
| D(allInclusive) | Indicator variable on whether mealtype is 'all inclusive' or 'full-board' (1 in both cases) or not (0) | Categorical |
| D(breakfastOnly) | Indicator variable on whether mealtype includes breakfast only or not (1 or 0) | Categorical |
| D(isHoliday) | Easter, Pentecost or Christmas within the holiday (1 or 0) | Categorical |
| weekday | Day of departure date (Monday, …, Saturday, Sunday) | Categorical |

**Table A.3:** Categorisation of the variable 'roomCategory'

| Indicator variable | Double room | High standard | Low standard | Sea view |
|---|---|---|---|---|
| Text string | 2-zimmer | deluxe | spar | meers |
| | 2 zimmer | superior | eco | mb |
| | dz | penth | | meerb |
| | 2 raum | villa | | sea view |
| | 2 räume | | | seaview |
| | doppel | | | meer-u |
| | zweizimmer | | | |
| | zweibett | | | |
| | double room | | | |
| | doubleroom | | | |
| | 2er | | | |
| | 2 be | | | |

Note: the indicator variable equals 1 if the variable roomCategory (converted into lowercase letters) contains one of the pre-defined text strings, and 0 otherwise. The text strings are defined according to the most frequent entries (top 100 values).

## A.2. Hedonic regression models: stability of coefficients and goodness of fit

As a necessity to the hedonic regression models in Section 4.2, the resulting coefficients have to be stable and plausible from an economic perspective. Coefficients of the *double imputation* model for each of the six holiday destinations are shown in Figure A.1. On the left-hand side, there are the coefficients for the variables ***travellerCount***, ***duration***, ***bookTime***, ***D_online*** and ***isHoliday***. As expected, all coefficients are positive, in other words the price of a package holidays increases with the number of travellers, the duration, the number of days the package has been booked in advance and if the holiday covers a period including one (or more) public holidays. One exception is for online bookings, signalling that a package holiday booked online is on average 8.4-11.9 % cheaper (depending on the holiday destination) than a package holiday booked offline via a traditional, high street travel agency. Concerning volatility over time, it has to be kept in mind that package holidays have a seasonal pattern, which will be reflected in volatile coefficients and partly also in a seasonal pattern ([55]).

The right-hand side of Figure A.1 shows the coefficients for the accommodation category of the underlying hotel, as indicated by one up to five stars. The benchmark in the regression model (3) is a four-star hotel, so five-star hotels are on average expected to be more expensive, whereas one- to three-star hotels are expected to be cheaper. This condition is fulfilled for nearly all holiday destinations. Besides this, the coefficient of a three-star hotel should on average be higher than for a two-star hotel, and the coefficient of a two-star hotel higher than for a one-star hotel. For most holiday destinations, this is true, but one- and two-star hotels are not common for all holiday destinations and therefore have only a small number of observations. This is reflected in the coefficients of one-star hotels, which are not stable for the Canary Islands and Turkey; for some months, these are higher than for two-star hotels or even positive, and they also exhibit missing values. The same problem occurs for two-star hotels in Egypt and the Dominican Republic. For example, the standard deviation of the coefficient for a two-star hotel in Egypt is higher ($\sigma = 0.09$) than for a three-star hotel ($\sigma = 0.02$) or a five-star hotel ($\sigma = 0.05$). The volatility of some regression coefficients (for example two-star hotels in Egypt) has only a minor effect on the index, because its implicit weight is very small. Concerning regular statistical production, the hedonic regression model could be adapted and optimised for each holiday destination. Nevertheless, most of the coefficients are stable and show a similar seasonal pattern.

([55]) For Greece and Turkey, note that the magnitude of the coefficients on ***travellerCount*** and ***duration*** exhibit a negative correlation during the summer months. Evidently, demand (also from non-German travellers) during this period is higher for these holiday destinations, which seems to rebalance the pricing scheme of tour operators concerning an extra day of stay and the number of travellers for the package holiday.

**Figure A.1:** Stability of regression coefficients over time (*double imputation method*)

**Figure A.1:** Stability of regression coefficients over time (*double imputation method*)

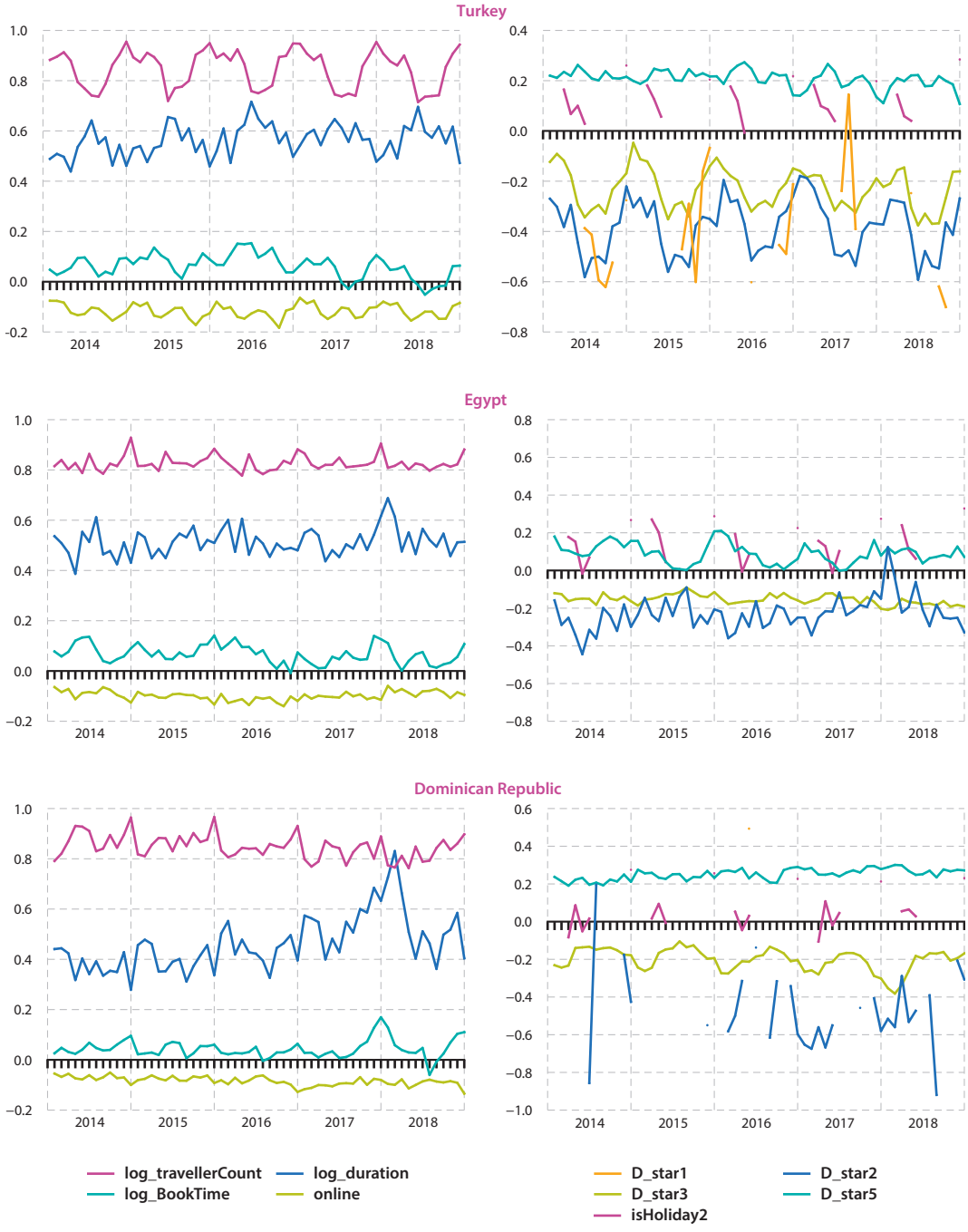


**Table A.4:** Adjusted $R^2$ by holiday destination

| Region\Method | Double imputation | | Time dummy model | |
|---|---|---|---|---|
| | Mean | Max-min range | Mean | Max-min range |
| Balearic Islands | 0.769 | 0.161 | 0.730 | 0.206 |
| Canary Islands | 0.721 | 0.113 | 0.677 | 0.118 |
| Greece | 0.753 | 0.118 | 0.695 | 0.156 |
| Turkey | 0.772 | 0.147 | 0.794 | 0.092 |
| Egypt | 0.704 | 0.205 | 0.717 | 0.175 |
| Dominican Republic | 0.785 | 0.121 | 0.720 | 0.099 |

## A.3. Detailed results for product definition following Chessa (2019)

**Table A.5:** Top-ten results of MARS for product definition of package holidays

| Number of combination | topArea | star | channel | bookTime_Class | depAirport | weekday |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | 1 | 1 | | 1 |
| 3 | 1 | 1 | 1 | 1 | 1 | |
| 4 | 1 | 1 | | 1 | 1 | 1 |
| 5 | 1 | 1 | 1 | | 1 | 1 |
| 6 | 1 | 1 | 1 | | | |
| 7 | 1 | 1 | 1 | | | 1 |
| 8 | 1 | 1 | 1 | 1 | | |
| 9 | 1 | 1 | | 1 | | 1 |
| 10 | 1 | 1 | | | 1 | 1 |

| Number of combination | Number of products | Mean of items per product | MARS | Homogeneity | Continuity |
|---|---|---|---|---|---|
| 1 | 10 681 | 193.2 | 0.33 | 0.40 | 0.79 |
| 2 | 1 008 | 2 047.5 | 0.33 | 0.35 | 0.93 |
| 3 | 1 582 | 1 304.6 | 0.32 | 0.35 | 0.92 |
| 4 | 5 423 | 380.6 | 0.32 | 0.38 | 0.83 |
| 5 | 2 752 | 750.0 | 0.32 | 0.36 | 0.88 |
| 6 | 36 | 57 330.1 | 0.32 | 0.32 | 1.00 |
| 7 | 252 | 8 190.0 | 0.32 | 0.33 | 0.95 |
| 8 | 144 | 14 332.5 | 0.32 | 0.33 | 0.95 |
| 9 | 504 | 4 095.0 | 0.32 | 0.33 | 0.93 |
| 10 | 1 379 | 1 496.7 | 0.31 | 0.34 | 0.90 |

Note: this table shows the top 10 results from MARS following Chessa (2019). The values of MARS are calculated as the average of 12 monthly MARS values in 2015. Combination No. 8 (highlighted in green) was taken for the main analysis in this paper which has a high mean of items per product, suggesting it has enough price representatives for a bias-free index calculation.

## A.4. Robustness of data filters and model specification

**Table A.6:** Construction of datasets (R1-R4) for robustness analysis

| Data filters used\Dataset | R1 | R2 | R3 | R4 |
|---|---|---|---|---|
| Excluding outliers as defined by the price per person per day and duration | X | X | X | X |
| German departure airports only | | X | X | X |
| Travellers > 16 years | | | | X |
| Excluding last minute bookings (within 14 days before departure) | | X | X | X |
| Online transactions only | | | X | |

Note: R2 denotes the baseline dataset used in the main analysis of the paper. R3 (online bookings only) also includes a more detailed regression equation for *double imputation*, as shown in Equation (5).

**Table A.7:** Number of observations used

| Holiday destination | Dataset R2 | | | Dataset R3 (only online transactions) | | |
|---|---|---|---|---|---|---|
| | Unit value price | Double imputation/ time dummy variable | Traditional stratification/ GEKS | Double imputation (1) | Traditional stratification (2) | Traditional stratification (3) |
| | | | | mealType/ roomType | mealType/ roomType | (top 25 hotels) |
| Balearic Islands | 491 382 | 470 069 | 446 394 | 129 434 | 118 350 | 70 715 |
| Canary Islands | 482 836 | 465 688 | 441 382 | 138 697 | 127 716 | 124 569 |
| Greece | 245 870 | 233 191 | 220 939 | 77 680 | 71 179 | 38 445 |
| Turkey | 658 706 | 637 694 | 633 795 | 200 131 | 197 587 | 102 828 |
| Egypt | 282 563 | 267 814 | 267 588 | 94 386 | 94 203 | 122 220 |
| Dominican Republic | 50 190 | 47 802 | 47 801 | 14 210 | 14 209 | 32 737 |
| **Total** | 2 211 547 | 2 122 258 | 2 057 899 | 654 538 | 623 244 | 491 514 |

(1) *Double imputation* based on the more detailed regression model in Equation (5).
(2) *Traditional stratification* according to ***star, D(seaView), D(AllInclusive)*** and ***bookTime_Class*** (***M = 48 strata***).
(3) *Traditional stratification* based on the top 25 hotels in each destination (as measured by their revenue share in 2015) according to ***iffCode*** and ***bookTime_Class*** (***M = 200 strata***).

**Figure A.2: Comparison of annual rates of change for traditional stratification using different datasets**

(% change compared with same month of the previous year)


Note: see Table A.6 for a description of the different datasets. R2 denotes the baseline dataset used in the main analysis of this paper.

*Source:* Bundesbank calculations on the basis of booking data from Amadeus Leisure IT GmbH

**Figure A.3: Comparison of annual rates of change for double imputation using different datasets**

(% change compared with same month of the previous year)


Note: see Table A.6 for a description of the different datasets. R2 denotes the baseline dataset used in the main analysis of this paper. R3 (online data only) also includes a more detailed regression equation, as shown in Equation (5).

*Source:* Bundesbank calculations on the basis of booking data from Amadeus Leisure IT GmbH

**Figure A.4: Comparison of annual rates of change for different versions of traditional stratification**
(% change compared with same month of the previous year)


Note: the baseline index refers to traditional stratification as used in the main body of this paper. Moreover, two alternatives are shown: i) a stratification at the hotel level (by ***iffCode***, ***channel*** and ***bookTime_Class***, ***M = 200 strata***) based on the top 25 hotels in each holiday destination (measured by revenue share in 2015), and ii) a more detailed stratification by ***star***, ***D(seaView)***, ***D(AllInclusive)*** and ***bookTime_Class*** (***M = 48 strata***) for online bookings only.

*Source:* Bundesbank calculations on the basis of booking data from Amadeus Leisure IT GmbH

## Getting in touch with the EU

**In person**

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact

**On the phone or by email**

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696 or
- by email via: https://europa.eu/european-union/contact

## Finding information about the EU

**Online**

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu

**EU publications**

You can download or order free and priced EU publications from EU Bookshop at: https://publications.europa.eu/eubookshop. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact).

**EU law and related documents**

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex at: https://eur-lex.europa.eu

**Open data from the EU**

The EU Open Data Portal (https://data.europa.eu/euodp) provides access to datasets from the EU. Data can be downloaded and reused for free, both for commercial and non-commercial purposes.

Print: KS-GP-19-002-EN-C
PDF: KS-GP-19-002-EN-N

Publications Office
of the European Union