Quality Assurance of Administrative Data: Northern Ireland House Price Index

This publication documents the quality assurance assessment of data sources for Northern Ireland used in the production of the Northern Ireland House Price Index (NI HPI) and subsequently the UK House Price Index (UK HPI).

Northern Ireland House Price Index - Quality Assurance of Administrative Data
Sales data from Stamp Duty Land Tax returns (HM Revenues & Customs)

Contents

1. Introduction

The Northern Ireland House Price Index (NI HPI) measures the change in the price paid to purchase residential property in Northern Ireland. The NI HPI is the NI component of the UK House Price Index, which is the single official measure of change in the price paid for residential properties in the UK.

The Office for National Statistics (ONS) take the raw data on prices paid for properties in England, Scotland and Wales to produce a GBHPI. Due to data restrictions, the raw data on prices paid for properties in NI cannot be transferred to ONS, so Land & Property Services (LPS) produces the NI HPI using the single agreed methodology and transfers the aggregate results to ONS. ONS then combine the GB index with the NI index to produce the UK HPI.

As the NI HPI in its entirety is used in the UK HPI, this assessment therefore serves as an assessment of the data for both the NI HPI and the UK HPI.

A number of different administrative datasets are used in the production of the quarterly HPI using a technique known as hedonic regression. In simple terms, hedonic regression is a technique which accounts for the changing quality of property transacted each period to isolate only pure price change, so that the change in price is not distorted by differences in the composition of property sold (for example, you can't directly compare the price of a one bedroom property sold in one period with a three bedroom property sold in another).

The hedonic regression approach requires detailed information on the characteristics of property sold, both regarding the physical attributes of the property (such as type, floor space etc.) and the location of the property (what type of neighbourhood, where in the country etc.). For the production of the NI HPI this data is obtained from a variety of administrative data sources that cover the price paid for transacted property (as reported to HM Revenues & Customs via Stamp Duty Land Tax returns), the attributes of a property (as held in the Northern Ireland Valuation List maintained by LPS) and characteristics related to the location of the property (such as the type of neighbourhood where the property is situated, defined by the ACORN classification from CACI Ltd). The quality of the Northern Ireland Valuation List database and ACORN classification data has been assessed separately and details of all of the assessments are available along with a short document summarising the overall quality of the data sources for the NI HPI.

This document will focus on the sale details of a property from the HMRC Stamp Duty Land Tax (SDLT) database, which is maintained by HM Revenues & Customs.

The majority of UK property transactions are liable for payment of Stamp Duty Land Tax. While the purchaser is liable for notification and payment of SDLT, completion of the land tax returns is usually carried out by the purchaser's solicitors on their behalf. The SDLT return and tax payment must be with HMRC no later than 30 days after the effective date of the property transaction. This should be the date when monies change hands.

The tax is paid on the value of the property and there are different rates of tax for residential and non-residential transactions. There are a number of transactions which are exempt from HMRC. These are:

- Property transactions where no money changes hands
- Property left in a will
- Divorce or dissolution of a civil partnership

Information contained within the Land Transaction Returns which are used by solicitors to notify HMRC of the details of taxable transactions, are provided to the Commissioner of Valuation in Northern Ireland under The Stamp Duty Land Tax (Use of Information Contained in Land Transaction Returns) Regulations 2009.

Regulation 3 of the 2009 Regulations states:

"Relevant information [meaning that on Land Transaction Returns] may be made available for use by the Commissioner of Valuation for Northern Ireland and District Valuers in Northern Ireland for the purposes of, the exercise of any statutory function of the Commissioner or District Valuers (as the case may be)."

Regulation 4 (1) of the 2009 Regulations states:

"Relevant information may be available for use by Department of Finance (DoF) for the purposes of any lawful function of DoF but must not be used in any way which would permit any person other than an officer of DoF to identify the vendor or the purchaser."

Publication of average transaction prices and use of transaction data in published research is permitted. However, care needs to be taken to avoid publishing data for an area that comprises:

- A single property of the type sold or
- A single transaction in a survey period

These circumstances could enable the specific property and thereby a vendor or purchaser to be identified.

The Stamp Duty Land Tax information is transferred securely from HMRC to LPS on a weekly basis. Each week sales not previously reported to LPS are extracted by an LPS statistician and processed. Address fields on the HMRC system are not validated, so LPS statisticians use address formatting software, to re-format the address of the property sold. The sales are then matched to the most recent monthly extract of the NI Valuation List, using address fields as the match keys, with the resultant data then used in the production of the latest quarter's house price index.

2. Risk rationale

The production and publication of house price data can be considered as **medium** profile, in that there is wider user/media interest in the results that are published, with

moderate economic or political sensitivity. Errors in the construction and publication of the statistics would have a moderate impact on the economy and affect the reputation of those responsible for the data.

The data quality concern attached to the sales data from SDLT returns used in the calculation of the NI HPI is considered a **medium** quality concern. This decision is based on the extent of the contribution of the small number of key fields (sale price and sale date) from SDLT returns which are used, in conjunction with the characteristics data from the NI Valuation List, in the calculation of the NI HPI. The NI Valuation List data has a much lower quality concern as the suppliers are based within LPS and the data collection and validation processes are well known. As the majority of data used in the NI HPI calculation is sourced from the NI Valuation List, this along with the validation routines carried out by LPS on the SDLT data, moderates the high risk factors from the small number of SDLT variables used in the calculations.

If SDLT data was the sole source of the NI HPI then it would be of **high** quality concern. This is because the data is collected by HMRC, via self-assessment, from multiple different sources (solicitors and house buyers) across the UK for tax purposes, it is assumed there is no co-ordinated training given to solicitors and buyers before completing the Land Transaction returns, and communication from suppliers is irregular.

Online guidance is available for solicitors/agents and HMRC have a number of automated validation checks in place as data is entered onto the system to safeguard the quality of the information. However, fields such as address of the property are not validated.

HMRC statisticians use a version of the data (accessed via a corporate data warehouse) as the basis for a number of National Statistics reports published on GOV.uk website, however LPS have no information regarding the transfer process from the live system to the corporate data warehouse. LPS will continue to carry out analysis on the coherence between statistics published by HMRC on sales volumes and those published in the NI HPI report and will update this document and related quality documents as necessary. As the dataset used by the HMRC statisticians is not a source of data for the NI HPI, the quality of this dataset is outside the scope of this report.

Further detail is provided in each of the areas below to explain the quality assessment.

When taking into consideration the public profile of house price statistics, its potential impact and the level of quality concern from the provider, the level of assurance attached to the use of the sales data from Stamp Duty Land Tax returns in the production of the NI House Price Index has been assessed as **A2: Enhanced assurance**.

As such, this Administrative data is deemed as being of medium risk, and an enhanced level of quality assurance is required.

Data Type	Administrative Source	Data Quality Concern	Public Interest	Matrix Classification
Sales Transaction Information	Stamp Duty Land Tax returns from HM Revenues & Customs	Medium	Medium	A2

3. Assessment of Stamp Duty Land Tax data using the Administrative Data Quality Assurance Toolkit

Land & Property Services has reviewed the Stamp Duty Land Tax return data against the Administrative Data Toolkit. The evidence meets the A2: Enhanced assurance category with regards to the sales data from the Stamp Duty Land Tax returns which is used in the production of the NI HPI.

3.1 Practice area 1: operational context and administrative data collection:

The Stamp Duty Land Tax return data is collected via self-assessment return(s) for the assessment and collection of tax. Some information is collected for the Valuation Office Agency for rating purposes. Currently, taxpayers are legally required to submit SDLT return(s) and payment to HMRC either digitally or by paper return, which are fed into a database of such returns. The main form used to collect this information is the SDLT1. HM Revenues & Customs publish a guidance manual for tax professionals which gives guidance on the forms and the content which should be provided. HMRC Stamp duty land tax manual

Digital returns are submitted either via the HMRC online system or via third party vendor (TPV) software which links directly into HMRC systems. The vast majority (approximately 97%) of returns are submitted digitally. Digital systems have in built data entry validation routines which check key fields such as local authority codes, and tax payer names are verified against other internal HMRC systems.

There are no known targets within HMRC which would introduce bias into the data collection process. Digital returns take minutes and paper returns are so few that they are generally scanned/entered on the same day, although payment issues may delay processing. Human error by the taxpayer or the agent is the most common issue. Usually, this is in the form of decimal point errors and omissions, geographic identifier entry errors, or relief claim errors. The error rate is much higher for paper returns.

Paper returns are keyed onto the system and are manually checked for clarity of text. Data entry validation, similar to that for digital returns, is used when operational

staff key information from paper returns onto the system. Certain variables, on both digital and paper returns, have restrictions on what can be entered (eg numeric, date). Some fields are mandatory, such as the taxpayer identifier and the unique transaction reference, on the online return the user cannot submit the return unless these fields are completed. If the return is rejected the customer is directed to complete an SDLT8 form online; on the paper return, if these fields are not completed, the return will be rejected and an SDLT8 form is sent to the customer for completion. Once the SDLT8 form is returned the correct information is entered into the system.

LPS receive a feed of “raw” data from the Stamp Duty Land Tax returns live database on a weekly basis. There are no compliance or statistical checks carried out, by HMRC, on the database before transfer to LPS. LPS have developed a suite of validation routines to select all residential sales records and verify them against the NI Valuation List.

3.2 Practice area 2: Communication with data supply partners

HMRC supply LPS with information from Stamp Duty Land Tax returns under the Finance No2 Act 2005 and subject to The Stamp Duty Land Tax (Use of information contained in land transaction returns) Regulations 2009.

Statisticians have met briefly with HMRC system suppliers to discuss the collection and validation processes of the SDLT data and to understand any issues around the data quality. LPS have gained a high level overview of the data entry validation processes used in the HMRC systems, and hopefully through further discussion LPS will develop a detailed knowledge of the data quality.

The data provided includes individual records of property transactions in NI, including the sale date, sale price, address of the property, names of the vendor and purchaser and the solicitor details for both parties.

The data are held securely within the LPS servers and are accessed by LPS statisticians using a secure electronic connection via the statistical package SPSS. The information contained in the dataset is sensitive as standalone data as it contains the names and addresses of the vendor and purchaser and also the address of the property sold. LPS statisticians, along with all LPS staff, are trained regularly in information assurance and adhere to the data protection principles to ensure that all data is protected appropriately and only non-disclosive data is released or published.

Based on a discussion with HMRC, LPS statistics branch accept the SDLT data from HMRC as accurate and assume that the information provided is of sufficient quality to collect tax and to use as a record of property transactions in NI.

3.3 Practice area 3: Quality Assurance principles, standards and checks applied by data suppliers:

LPS receive an extract of the “raw” data from the Stamp Duty Land Tax database at the beginning of each working week. The data shows all transactions up to the day before the extract date and is used as the census of sales to feed the calculation of the quarterly NI HPI.

The sales database contains information on the price paid for the property, the date of sale, the address of the property sold, and the names and addresses of the vendor, purchaser and their agent(s).

HMRC have supplied SDLT data in an electronic format to LPS since 2005, prior to this sales details were provided in hard copy on Particulars Delivered forms (PD1). The calculation of the NI HPI only makes use of sales provided electronically and the series runs from 2005 to present. It is not possible to analyse sales reported on PD1 forms prior to 2005.

The taxpayer is responsible for entries on the return, both paper and digital, which feed the database. HMRC staff are responsible for checking that mandatory fields such as unique transaction reference number, National Insurance Number and Local Authority Codes are complete. If the information for some fields (stated in the digital form) for example no consideration entered, is incomplete, HMRC will write to taxpayer/agent to obtain correct/full information before entering the sale onto the system. Paper returns will be processed if all the mandatory fields are completed.

There are no checks carried out on property or tax payer address fields entered or typing errors for the raw data which is transferred to LPS. LPS only use sales where there is a valid address which can be matched to a property in the NI Valuation List. Therefore LPS statisticians validate the data from HMRC against the Royal Mail Postal Address File via address matching software and manually check sales which fail to match a Royal Mail address. Around 25% of residential addresses supplied on SDLT returns require manual checking and matching to a property in the NI Valuation List.

Each sales record is uniquely identified by a Document Number and a unique transaction reference number, however it is still possible for duplicate records to occur in the dataset, some where the unique transaction reference number is duplicated. HMRC have suggested that duplicate records are caused by agents submitting a return twice. This can occur for a number of reasons eg to correct an error on the first form, two staff members dealing with the same case and both submit a return, or in some cases the agent believes SDLT did not receive the return and so another one is submitted. LPS will continue to identify and remove duplicates from the NI House Price Index calculations. HMRC may remove duplicate cases at a later date if taxpayers notify them when they receive duplicate correspondence (eg requests for payment)

There is medium risk to the quality of the administrative data collected for the Stamp Duty Land Tax database as, although there is a legislative reason underpinning collection of the data which raises money for Government services, there are limited

checks and validation carried out on the data entered into the SDLT database. The collection of data via digital returns has improved standardisation, and subsequently the quality of information received by HMRC, as there are a number of in-built validation checks which must be met before the return is accepted (for example certain mandatory fields cannot be left blank and dates must be in the correct format).

3.4 Practice area 4: Producers Quality Assurance Investigations and documentation

The SDLT data is a set of property sales that is used as input in the production of modelled house prices, which in turn are used in the production of the quarterly NI House Price Index and also the monthly UK House Price Index (HPI). Read the full details on the NI HPI production methodology

On receipt of the SDLT data LPS statisticians carry out a number of high level validation checks:

- Select all sales which are residential (Property Type is 01 & 04)
- Flag sales with a zero or missing sale price
- Flag land sales using fuzzy techniques on free text address fields
- Manually scan sales for multiple properties & poor addresses which can never identify an individual property
- Remove sales of land, multiple properties, those with sale price zero and poor addresses
- Format remaining sales addresses to Royal Mail Postal Address File (PAF) format using address matching software
- Identify and remove duplicate sales
- Flag non market value and co-ownership sales

There are issues with missing data in key fields within the SDLT database for example building number, street or postcode. LPS statisticians use all of the information available to match as many sales as possible to the NI Valuation List, to ensure the NI HPI is based on the best sales evidence possible.

LPS statisticians carry out further quality assurance on a weekly basis to ensure the matching of sales data to the NI Valuation List takes place successfully, and the resulting modelled house price data is of sufficient quality. Transactions with extreme price values are checked against the capital value held within the Valuation List and/or the sales brochure where available. Any sales which cannot be verified are excluded from the analysis. The modelling process used in the production of house price data includes an automated statistical technique (Cook's Distance) that assesses modelled house prices for property with a certain set of attributes (which are derived using the various sources of input data) against the price for a similar property. If the modelled price is substantially different (meaning it exceeds a predefined tolerance) then the price is excluded from being used in the final house price estimate.

4. Strengths and limitations of data

The returns provided by solicitors for Stamp Duty Land Tax are the only available source of information for sales of properties in Northern Ireland. HMRC supply LPS with information from Stamp Duty Land Tax returns under the Finance No2 Act 2005 and subject to The Stamp Duty Land Tax (Use of information contained in land transaction returns) Regulations 2009.

It is clear from the data that solicitors/agents notify HMRC of both liable and non-liable transactions, therefore the majority of sales are included in the dataset. More than 90% of sales, which have enough address information to identify a specific property, can be verified against a property in the NI Valuation List.

However, there are some acknowledged limitations in the data:

Considerable reliance is placed on agents to enter a full correct address on the SDLT return as the address fields are not validated on entry to the system. LPS have developed a suite of detailed validation routines which re-format the address, remove sales of multiple properties recorded under a total sale price and land sales, to match as many sales records as possible to the appropriate residential properties the NI Valuation List. There will always be a small number of sales records (approximately 10%) which, due to poor/incomplete address details, cannot be verified as a residential property in Northern Ireland and are therefore excluded from any price calculations.

There are a number of duplicate cases present within the data, however these are easily identifiable and are removed before any analysis is undertaken.

Overall, LPS are confident that the data received from HMRC is of good quality to produce reliable results for the NI HPI (and subsequently the UK HPI). The reliability of the results have been confirmed by sense checking against other house price measures and through feedback from economists, trusted for their opinions on the property market.