Exploring clusters in social housing data

Vicki Howe
Dr Grazziela Figueredo
Dr Philip Quinlan
Dr Duncan R Shaw
<table>
<thead>
<tr>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Foreword</strong></td>
</tr>
<tr>
<td><strong>Executive Summary</strong></td>
</tr>
<tr>
<td><strong>Introduction</strong></td>
</tr>
<tr>
<td><strong>Collecting the data</strong></td>
</tr>
<tr>
<td>Stages of data collection</td>
</tr>
<tr>
<td>Types of data and standardisation</td>
</tr>
<tr>
<td>Legal and ethical issues</td>
</tr>
<tr>
<td><strong>Building the database</strong></td>
</tr>
<tr>
<td>Construction of the database</td>
</tr>
<tr>
<td>Deprivation data</td>
</tr>
<tr>
<td>Crime data</td>
</tr>
<tr>
<td><strong>Cluster analysis</strong></td>
</tr>
<tr>
<td>Purpose of cluster analysis</td>
</tr>
<tr>
<td>Consensus clustering</td>
</tr>
<tr>
<td>How many clusters?</td>
</tr>
<tr>
<td><strong>Results of cluster analysis</strong></td>
</tr>
<tr>
<td>Clusters</td>
</tr>
<tr>
<td>Income management and the clusters</td>
</tr>
<tr>
<td>Practical application of clusters for income management</td>
</tr>
<tr>
<td>Repairs hotspots</td>
</tr>
<tr>
<td>Practical application of clusters repairs</td>
</tr>
<tr>
<td><strong>Discussion</strong></td>
</tr>
<tr>
<td>Results from the pilot project</td>
</tr>
<tr>
<td>Encouraging participation</td>
</tr>
<tr>
<td>Specification of data</td>
</tr>
<tr>
<td>Collection of data</td>
</tr>
<tr>
<td>Integration and validation of data</td>
</tr>
<tr>
<td>External data</td>
</tr>
<tr>
<td>Analysis</td>
</tr>
<tr>
<td>Interpretation</td>
</tr>
<tr>
<td>Bonus side-effects</td>
</tr>
<tr>
<td><strong>Explore your own data</strong></td>
</tr>
<tr>
<td>How HouseMark can help</td>
</tr>
<tr>
<td><strong>Appendix</strong></td>
</tr>
<tr>
<td>Appendix A: Property locations - list of local authorities</td>
</tr>
<tr>
<td>Appendix B: List of landlord data collected</td>
</tr>
<tr>
<td>Appendix C: List of data used in clustering</td>
</tr>
<tr>
<td>Appendix D: Acknowledgements</td>
</tr>
</tbody>
</table>
There is growing interest in the social housing sector in learning more about customers and properties using ‘data mining’ techniques applied to combinations of data held by landlords and external agencies.

There have been several attempts in social housing to discover actionable insight from complex Big Datasets which have failed – largely due to the data collection inexperience of the project sponsors and poor project design and consequent limitations on data integrity.

This report sets out the findings of what we believe to be the first successful attempt to ascertain practical business intelligence by developing illustrative clusters of meaningful data from a mass of complex information.

We believe that the insight in this report stems from the care and preparation that underpins the approach taken by HouseMark, our data scientist partner ADAC, Nottingham University Business School and in particular the landlords which participated in the ‘discovery pilot’ study that forms the basis of this report.

This report shows what is possible – focusing particularly on repairs hotspots and aspects of income management. It demonstrates the benefits of landlords working together to collect and analyse data at a local level to tackle issues of common interest.

The report should be read in conjunction with the first comprehensive guide to data protection in the social housing sector – ‘Transparency & Trust’ - which was jointly published by HouseMark, Anthony Collins (solicitors) and AmicusHorizon in January 2017.

For landlords, whom having read this report would like to commence data mining projects, HouseMark is now able to provide practical consultancy and project management support through our working relationships with ADAC, Nottingham University Business School and Anthony Collins (Solicitors).

We believe that this report is groundbreaking in our sector and we welcome reader comments, suggestions and challenges.

Arturo Dell
Director of Product Development,
HouseMark
Why this project?
There have been few practical demonstrations of the capacity of Big Data techniques to transform our understanding of how the complex factors that affect social housing businesses, our properties and our tenants interact.

HouseMark decided to get stuck in and work with our member landlords and our partners at Nottingham University in this discovery project. Our aim was to check out the real world potential to use Big Data techniques to look for patterns in a large set of property and tenancy-related data from a number of landlords, combined with public open data.

Who is this project for?
This report has been produced for the eight landlords (councils and housing associations) participating in the study and for all those interested in using the latest data analytical techniques to investigate their housing data. It describes how we tackled the practical issues around data collection, consistency and privacy, and demonstrates the potential of data clustering techniques to deliver useful new insights not apparent from the raw data.

What does this analysis tell us?
Our analysis identified that two out of our seven clusters of properties had a significantly higher likelihood of generating high numbers of repairs, even after the data had been ‘normalised’ to take account of the correlations between number of repairs and bed size of property.

It also showed the influence of property type in one of the high repair groupings, where it was clear that the ‘hotspots’ were concentrated in flats within that cluster.

The analysis also found that the two clusters which had much higher rates of the tenants making their rental payments by cash, had significantly less legal action for arrears.

Perhaps as important as the specific findings, we have shown the potential that cluster analysis of large data sets has to deliver useful new insights not apparent in the raw data.

How can this information be used?
The ability to identify groups of homes with very different repairs load and legal arrears activity can be used to re-invent landlords’ approaches to housing management by targeting communications, policy and interventions.

We have also learned valuable lessons about how to make a collaborative Big Data project work in practice. We drew out a number of learning points through the report that will be valuable to others following this path.

Applying these techniques to other datasets
This was a discovery project – a ‘proof of concept’. The methods used here can be extended to other data to explore a wide variety of other questions, such as income management, tenancy sustainment or community cohesion.

There is also scope for extending the range of public open data used, e.g. census data, educational attainment, and economic activity.

Register interest for future projects
If you’re interested in taking part in future projects with HouseMark and our partners, please contact Nayna Kavia via email at nayna.kavia@housemark.co.uk.

HouseMark services
Readers may be interested in commissioning HouseMark to apply Big Data techniques to their own ideas for areas of enquiry. In association with our data scientist colleagues from the Advanced Data Analytics Centre of Nottingham University, we can provide consultancy support to assist with analysis and project management, as well as identifying data sources and funding opportunities, and providing data protection assurance.

To find out more on how HouseMark can help you, please get in touch with us on 024 7647 2703 or email consultancy@housemark.co.uk.
The challenge of Big Data

‘Big Data’ refers to the collection, management and analysis of large data sets to improve business processes. It is about helping an organisation to better understand its customers and hence to provide improved or more cost-effective services. Big Data is typified by the use of ‘data mining’ techniques. Data mining involves the extraction of patterns and ‘knowledge’ from large aggregations of data.

Big Data and housing

Social landlords hold a wealth of data about their properties and tenants. For each, this is an asset which could be exploited to identify patterns. When the data from several – or many – landlords is combined, it becomes possible to identify deeper and more subtle patterns and processes. Using advanced computational techniques, this data can be combined with data from governmental sources, such as the Office for National Statistics, or commercial sources, adding context and increasing its value. Emerging patterns and associations can help target certain tenancy interventions and predict likely future outcomes for spend and performance.

The pilot project

In 2014 HouseMark began working with a group of its landlord members to explore the potential for using Big Data approaches to investigate patterns in large amounts of property and tenancy-related social housing data in combination with public open data.

This ‘discovery project’ focused on developing robust data structures and definitions that would ensure consistency and extensibility in a large data set gathered from multiple sources.

Preliminary analysis then examined whether clustering techniques could reveal hitherto unrecognized groupings of similar entities within the data.

Finally, the potential usefulness of the clustering to develop predictive models was demonstrated using the example of repair hotspots within social housing stock.

Collaborators

The project was a joint enterprise between HouseMark, Nottingham University’s Advanced Data Analysis Centre (ADAC) and Nottingham University Business School (NUBS). HouseMark provided the coordination and data collection from pilot landlords while ADAC performed the data integration and analysis.

HouseMark is the UK’s leading provider of social housing data and insight, seeking to drive improvement in the sector by giving its members the tools they need to respond to change. HouseMark has around 950 housing association, local authority and ALMO members across the UK.

ADAC is a data analysis team from the University of Nottingham that provides leading data analytical procedures to real-world challenges.

NUBS is a leading centre of research into Qualitative Data Science, which uses ideas from complex systems to suggest data monetisation strategies and digital services.

The data itself was provided by eight member organisations of HouseMark: seven housing associations and one council landlord.

Collaborators

1 Extensibility is a systems design principle where the implementation takes future growth into consideration. It is a systemic measure of the ability to extend a system and the level of effort required to implement the extension.

<table>
<thead>
<tr>
<th>Landlord name</th>
<th>No. of properties in study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspire Housing</td>
<td>8,796</td>
</tr>
<tr>
<td>Devon and Cornwall Housing</td>
<td>19,725</td>
</tr>
<tr>
<td>Leeds Federated Housing</td>
<td>3,363</td>
</tr>
<tr>
<td>City of Lincoln Council</td>
<td>7,968</td>
</tr>
<tr>
<td>mhs homes</td>
<td>7,648</td>
</tr>
<tr>
<td>Orbit Housing</td>
<td>35,660</td>
</tr>
<tr>
<td>South Liverpool Homes</td>
<td>3,616</td>
</tr>
<tr>
<td>Trafford Housing Trust</td>
<td>9,223</td>
</tr>
</tbody>
</table>

Table 1 - Pilot landlords and their stock

Together, these organisations supplied data for over 95,000 properties of all sizes, located across 53 English local authorities (listed in Appendix A). Their cooperation was crucial to the success of this pilot project.
Stages of data collection

The discovery project had three stages of data collection which are summarised in Table 2. All the data related to 31 March 2014.

### Stage 1 data

The Stage 1 data collection was the model for all stages. Data was classified as ‘essential’ or ‘desirable’. If a landlord could generate the desirable data with minimal effort, it was requested to do so, but not to go to extraordinary efforts.

Data was restricted to residential properties that are both owned and managed by the landlord. For example, garages and shops were excluded, as were properties that were either owned or managed by third party organisations.

Social landlords have statutory and regulatory reporting requirements which lead naturally to common standards for some data. The definitions for this project were mostly based on the Homes and Communities Agency’s (HCA) NROSH+ Guidance Notes for the Statistical Data Return and the Department for Communities and Local Government’s (DCLG) National Register of Social Housing Field Definitions.

#### Stage 2 data

Stage 1 included both self-contained units and bedspaces in non-self-contained accommodation, while Stage 2 used only self-contained units, excluding non-self-contained housing.

#### Stage 3 data

Stage 3 was tenancy-related, so data was not collected for leaseholder or shared ownership properties or voids. Because of the legal and ethical need to ensure individuals could not be identified from the data, landlords aggregated information for Stage 3 based on geographical area and age band of the lead tenant.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Description</th>
<th>Examples of data</th>
<th>Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Basic property type and location data</td>
<td>Outward postcode, Number of bedrooms, Tenure type, Social/affordable/market rent, Number of repairs</td>
<td>8 landlords reporting 95,899 properties</td>
</tr>
<tr>
<td>2</td>
<td>More property attributes</td>
<td>Occupancy limit, Construction type, Floor area, Repairs costs, Existing use values</td>
<td>6 landlords reporting 83,060 properties</td>
</tr>
<tr>
<td>3</td>
<td>People data</td>
<td>Postcode sector, Arrears bands, Anti-social behaviour reports, Person age band</td>
<td>3 landlords reporting data for 52,104 properties in 3,151 postcode sectors</td>
</tr>
</tbody>
</table>

Table 2 - Stages of data collection - A full list of data supplied by the landlords is given in Appendix C, with the data included in the analysis detailed in Appendix D.
Learning point 1 - Getting started

Landlords hold a wealth of information on their systems about properties, tenancies and occupiers. It is important to set aside time at the beginning of a Big Data project to consider:

- what data sources are available
- how the data can be extracted and formatted consistently
- how up-to-date and accurate each data item is
- to develop and document clear and unambiguous data definitions and field names into a data dictionary.

Wherever possible, pre-existing definitions (e.g. those used by regulators or government agencies) should be used.

Types of data and standardisation

Each piece of data has:

- a description of what it represents
- a short 'field name' in the database
- a definition of the allowable values

Some data is naturally numeric, such as rental amount or number of bedrooms. Others, such as type of central heating, need a different approach. Detailed definitions for the data required at each stage were supplied to the collaborating organisations. Tables 3 and 4 show examples of these definitions.

Legal and ethical issues

Postcodes provide a very useful means for identifying groups of properties which can usefully – if not always accurately – be assumed to be one homogeneous area. Hence, a good approach for some applications is to use the postcode as the main geographical identifier.

However, there are legal and ethical issues in the collection and use of personal data which is transferred out of the landlord’s organisation.

Much of the data can be anonymised so that a third party could not track it back to individuals. This is the approach we took in this project.

We also elected to work with postcode sectors rather than full postcodes. A postcode sector comprises the first part of the postcode, the space and the digit immediately following the space. For example, HouseMark’s full postcode is ‘CV4 7HP’; the corresponding postcode sector is ‘CV4 7’.

Given a full postcode, a landlord and an unusual property type or size, a malicious person – known in ICO jargon as a ‘motivated intruder’ – might be able to identify the actual address and occupants.

In addition, in general terms, individual postcodes are often at too fine a resolution to be useful, but postcode sectors can offer the required granularity.

The approach complied with the Information Commissioner’s Office 2012 code of practice ‘Anonymisation managing data protection risk’.

Learning point 2 - Protecting individuals

Give early consideration to practical and ethical considerations around location data – what level of granularity is appropriate for your analysis and how can you prevent inadvertently allowing individuals to be identified from the findings of your analysis of their personal data?

HouseMark’s Guide to Data Protection and Privacy – entitled ‘Transparency and Trust’ – should be your starting point; HouseMark and its legal associates can provide further expert assistance if required.

Table 3 - Example of data definition: property construction

<table>
<thead>
<tr>
<th>Field name</th>
<th>Abbreviated field name</th>
<th>Values - full name</th>
<th>Values - code name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purpose built / converted</td>
<td>PB_CON</td>
<td>Purpose built PB</td>
<td></td>
</tr>
<tr>
<td>Converted</td>
<td>CON</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dwelling type</td>
<td>TYPE</td>
<td>Bungalow BLING</td>
<td></td>
</tr>
<tr>
<td>Dormer bungalow</td>
<td>DORM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>House</td>
<td>HOUS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flat</td>
<td>FLAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maisonette</td>
<td>MAIS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>OTH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form of structure within</td>
<td>FORM</td>
<td>Detached DET</td>
<td></td>
</tr>
<tr>
<td>which dwelling is located</td>
<td></td>
<td>Semi-detached SEMI</td>
<td></td>
</tr>
<tr>
<td>Terraced</td>
<td>TERR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>End-terrace</td>
<td>END_TERR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Back to back terrace</td>
<td>BACK_TERR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Block</td>
<td>BLOCK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>OTH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type of construction</td>
<td>CONST</td>
<td>Traditional TRAD</td>
<td></td>
</tr>
<tr>
<td>Non-traditional</td>
<td>NON_TRAD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-permanent</td>
<td>NON_PERM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4 - Example of data definition: property age and size

<table>
<thead>
<tr>
<th>Field name</th>
<th>Abbreviated field name</th>
<th>Value descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floor area</td>
<td>AREA</td>
<td>Numeric</td>
</tr>
<tr>
<td>Number of bedspaces</td>
<td>BSp_Count</td>
<td>1 to 99</td>
</tr>
<tr>
<td>Year built</td>
<td>YEAR</td>
<td>Year (YYYY)</td>
</tr>
</tbody>
</table>
Building the database

Construction of the database
The data was securely transferred to ADAC who constructed a relational database using MySQL.
The first phase of this was to adjust the data supplied by the landlord organisations in CSV and Excel files to ensure its compatibility with the required data definitions. In general, this was a simple operation.
The second phase was to load the landlords’ data from CSV or Excel files into database tables.
The third phase was to construct appropriate linkages and look-ups, such as those described in following sections.

Deprivation data
Relating deprivation data to properties
The English Indices of Deprivation measure relative levels of deprivation in small areas of England called Lower-layer Super Output Areas (LSOAs). The geographical boundaries of LSOAs are available from ONS, but these are not directly mapped to postcodes.
The solution to this mapping issue is typical of problems faced in Big Data projects. Quality of data must be balanced against the effort involved in refining data. The strategy here was based on postcode sectors.
The process used was:
1. get all postcodes for the given sector
2. get the populations for each postcode in that sector
3. select the postcode from that sector with the largest population
4. get the LSOA for the selected postcode
5. use that LSOA for the postcode sector

The result was that the deprivation data used was taken from the largest postcode from that given postcode sector. This is not perfect but is a reasonable estimation.

Representing deprivation levels of the clusters
In the English Indices of Deprivation 2010, the LSOA with a rank of 1 is the most deprived and the LSOA with a rank of 32,844 is the least deprived.
In this project, adjusted deprivation data is presented on a numerical scale – with the percentage for each cluster representing its position on that scale. The lower the rank, the higher the percentage and the greater the deprivation.

Learning Point 3 - Make findings understandable
You may need to transform some of your measures in order to make the findings understandable to your intended audience.
In this project, we ‘flipped’ the direction of the deprivation ranking from ‘1=highest deprivation’ to ‘1=lowest deprivation’ since it can be counter-intuitive for a low number to represent a high level of deprivation. The transformation made the data more meaningful and the findings easier to understand.

This method of selecting a single postcode for mapping to a single LSOA was chosen because it was straightforward to understand and sufficiently robust for this project.
Cluster analysis

Purpose of cluster analysis
Cluster analysis comprises a suite of techniques which attempt to identify groups of items (clusters) that are similar to each other, where each item is a member of exactly one cluster.

This study was designed to test whether clustering techniques could identify segments within our social housing dataset, and deliver useful insights to support decision making.

Consensus clustering
Different clustering algorithms can produce different results, so a prudent approach is to use several techniques and look for a consensus. ADAC’s data analysts used hierarchical clustering (HCA), k-means (KM) and Partitioning Around Medoids (PAM). These techniques were chosen as they are among the most widely used clustering methods in data mining.

How many clusters?
The best number of clusters is a tenuous concept. Clusters must be large enough to be useful, but small enough that they have well-defined characteristics.

The ‘best’ number of clusters was defined as the number that gave the most consensus between the deployed techniques. Six types of validity indices were applied and according to their specific rules, they indicated the appropriate number of clusters for this analysis.

Learning point 4 - Expert help when needed
Choosing how many clusters to investigate can affect your findings. Expert advice from experienced data scientists will help you choose the best number for your data.

HouseMark has developed a network of experts who can provide you with this support.
Results of cluster analysis

Clusters
The consensus clustering analysis identified seven clusters.
Cases that did not gain consensus were placed in an eighth cluster, containing 971 data points, which was not investigated further in this study.

Learning point 5 - Identifying ideas for further research
Further analysis of data that does not achieve consensus might reveal interesting sub-clusters. This discovery project was unable to investigate this data further, but it would be worth doing so in future studies.

The data fields that best differentiated the clusters were:
• payment methods
• arrears levels (<£500 and £500-£1,000)
• legal action taken (no legal action or NSP issued)
• ethnicity
• deprivation

These fields have been used to construct illustrative ‘dashboards’ for each cluster.

Learning point 6 - Clear presentation of results
The visual presentation of the findings and naming of clusters can be helpful in interpreting your findings and gaining greater insight into your data.

The symbols used in the descriptive dashboard graphics for each cluster indicate:

- Statistically significant highest value compared to other clusters
- Statistically significant high value compared to other clusters but not the highest
- Statistically significant lowest value compared to other clusters
- Statistically significant low value compared to other clusters but not the lowest

- Cash/cheque
- Direct debit
- Direct benefit
- Post Office
- Standing order
- Swipe card
Results of cluster analysis

Cluster 1: Small flat-dwellers in deprived neighbourhoods
- Highest deprivation
- Highest concentration of flat-dwellers
- About average on NOSPs
- Relatively high cash payers, low proportion on HB Direct

Cluster 2: Cash-economy, striving minority ethnic flat-dwellers
- By far the highest concentration of cash rent payers
- Second highest deprivation levels
- Few on HB Direct
- High non-White British
- Lowest level of arrears between £500 and £1,000
- Unlikely to have legal arrears action taken against them
Cluster 3: Benefit-dependent White British flat-dwellers
- High deprivation
- Almost exclusively living in flats or maisonettes
- Second highest concentration of White British
- Most likely to have legal arrears action
- Highest users of HB Direct and several other rent payment methods

Cluster 4: Struggling flat-dwellers in less deprived neighbourhoods
- High proportion of flat dwellers
- Low use of cash and swipe cards
- Low deprivation levels
- Second highest concentration of arrears between £500 and £1,000
Cluster 5: Struggling people in least deprived neighbourhoods
- Least deprivation levels
- Wide mix of property types
- Lowest level of cash payers and highest concentration of swipe card users
- Most likely to have arrears between £500 and £1,000
- Second highest level of NOSPs

Cluster 6: White British people in family homes
- Highest White British
- Middling deprivation
- Majority living in two and three bedroomed homes
- Most likely to pay by Direct Debit
- Second most likely to have HB Direct payments
- Middling arrears levels
- Second highest NOSP levels
Results of cluster analysis

Cluster 7: White British house-dwellers
- Joint highest White British
- By far the highest proportion of houses
- Most likely to pay by swipe card
- Second most likely to pay at Post Office
- Middling on arrears levels

Income management and the clusters

This pilot project aimed to explore whether clustering techniques applied to large databases would reveal interesting insights, and whether the clusters could inform predictive models that could be used in a practical way to inform decisions. ADAC’s data analysts and HouseMark first looked at inter-relationships between data fields that define the clusters. The strong relationship between rental payment method(s) and a propensity for legal arrears action was particularly striking.

The Cluster 2 grouping – cash economy striving minority ethnic flat-dwellers - exhibits the highest level of payment by cash or cheque, with 64% of members paying in this way. Direct Debit is used more frequently than by other clusters, while usage of direct housing benefit, Post Office orders, standing orders and swipe card payments are significantly lower than the other clusters.

Cluster 2 also has the highest proportion of members with arrears below £500 but the lowest proportion with rent debts between £500, and £1,000. This relatively low level of debt is reflected in the rent arrears actions most frequently found within the group: it is the cluster with the highest percentage of members who have had no legal action taken at 87%, and the cluster with the lowest percentage where a notice of seeking possession (NSP) has been issued at 8%.

The cluster is predominately not White British, with only 20% falling into this ethnic category, the lowest of any cluster. Cluster 2 represents reasonably deprived areas as they are higher than most clusters (although not the highest).

Table 5 – Arrears action and cash payments by cluster.
Results of cluster analysis

Practical application of clusters for income management

These findings might inform a landlord’s policy and practice in a number of ways, such as:

- supporting landlords in introducing a more preventative focus, rather than reactive, for income management activities such as debt counselling and benefit advice
- targeting (by cluster group) the promotion of money advice services and debt reduction strategies
- directing resources to specific customer groups to assist them in sustaining their tenancies
- identifying those residents that do not have access to bank accounts and pay by cash to provide support for them in managing their finances and for the receipt of any benefits
- identifying the most appropriate rent payment methods for individual tenants
- reducing the transaction costs a landlord incurs for receiving rental payments, by supporting tenants to move to more cost efficient methods of payment
- considering the options the landlord makes available for paying rent to suit identified groupings, and nudging behaviour towards the most cost efficient methods.

Repair hotspots

To further test the validity of the clusters, ADAC’s data analysts and HouseMark then added new data relating to repairs volume and looked at whether the clusters that a property belonged to could be used to predict the likelihood that it would experience a higher than normal volume of repairs.

Properties which have a high repair load were designated as ‘repair hotspots’. It might be expected that a large property with several residents would require more repairs than a smaller property occupied by one or two people, as the smaller home will have fewer radiators, fewer doors, fewer windows, etc.

Hence the working criterion was based on the average number of repairs for a property of a particular type and size. A property that required more than 150% of the mean number of repairs for its type and size was considered to be a repair hotspot. The values applied are listed in Table 6.

In brief, the analysis demonstrated that the prevalence of repair hotspots was greater in some clusters than others, and that the pattern is different if you look only at flats than if you look at all properties. This suggests that property type is an important variable in determining the likelihood of repair hotspots. The panel below discusses these findings in more detail.

<table>
<thead>
<tr>
<th>Type</th>
<th>Bedrooms</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bungalow</td>
<td>1</td>
<td>6.45</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6.06</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>6.65</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>9.92</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>7.88</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>9.00</td>
</tr>
<tr>
<td>Flat</td>
<td>1</td>
<td>5.98</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>4.72</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>5.48</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>6.10</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>7.00</td>
</tr>
<tr>
<td>House</td>
<td>1</td>
<td>5.03</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6.79</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>6.14</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>8.12</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>10.13</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>14.40</td>
</tr>
<tr>
<td>Maisonette</td>
<td>1</td>
<td>6.71</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6.14</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>6.35</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>7.00</td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
<td>3.96</td>
</tr>
</tbody>
</table>

Table 6 – Thresholds for repair hotspots

However, further investigation showed that property type is also associated with repair hotspots. Looking only at flats, the statistical significance of the association is greater than that seen across all properties (p=0.001). This phenomenon is particularly marked in Cluster 5 which is one of the groupings experiencing higher prevalence of repair hotspots amongst flats, whereas it was the lowest cluster for hotspots across all property types.

Figure 1 - Distribution of repair hotspots across clusters in all property types

Figure 2 - Distribution of repair hotspots across clusters in flats

An initial analysis indicated that there is a statistically significant association between the clusters and whether repair hotspots exist within these clusters (p=0.05). Clusters 2 and 7 have higher hotspots than the others.
Results of cluster analysis

Practical application of clusters for repairs

These findings might inform a landlord’s policy and practice in a number of ways, such as:

- identifying high service users so that underlying problems can be ascertained and issues addressed
- targeting the provision of:
  - pre-tenancy training
  - DIY training for existing tenants
  - repair incentive schemes
- reviewing:
  - component specifications
  - preventive maintenance practice for high repair properties
  - lettings policies for high repair properties.
- supporting active asset management strategies:
  - identifying which properties to sell rather than invest in
  - developing or acquiring stock with different characteristics from those of high-repair properties
- managing customer expectations. Some cluster groups may have different expectations on the repairs service than others. For example, some may expect a landlord to come out and clear a blocked sink, while others may fix it themselves
- profiling reactive and planned maintenance spend, and using clusters to estimate repair costs based on the landlord’s property types and household sizes over future years

Learning Point 7 - Proof of concept

Our investigation of the distribution of repair hotspots and income management interventions demonstrates that cluster analysis can highlight differences and relationships in social housing stock that are not apparent in the raw data and lead to a range of practical uses.

It opens the real possibility of using the technique on more comprehensive data sets from a single landlord or many different landlords (at a local level) to give valuable actionable insights that may improve the efficiency and effectiveness of the social housing sector.
Discussion

Results from this pilot project
This project has demonstrated that data does indeed fall into clusters which are capable of being analysed to produce actionable insights.

The following sections review the experience gained from this project and look forward to future developments.

Encouraging participation
More good quality data will give better results and allow us to perform deeper analyses.

The pilot project focused on two sample topics of interest, but wider participation would allow more topics to be investigated. HouseMark hopes to create a ‘virtuous circle’ where useful results from early projects generate interest, so more participants help produce larger and better data sets, further improving and extending results.

Specification of data
HouseMark used its experience of data collection and practical housing issues to produce an initial data specification. This was revised in the light of feedback from participants so the final specification was clear and straightforward to implement.

Where possible, data fields were based on quantities that are routinely used in housing management, especially those which have tight definitions for reporting to government and regulators.

Deeper, wider investigations will allow us to collect more types of data. Specifying these so they can be collected efficiently and accurately requires care.

As the scope of analyses widens, considerable attention is needed to avoid conflict with ethical and legal principles. This can sometimes be achieved by anonymising data or by aggregating it so no individual tenant can be identified by anybody outside the landlord’s organisation.

HouseMark will review these issues as they arise and construct paths that provide good data, ethically and legally sourced. HouseMark’s ‘Transparency and Trust – a Guide to Data Protection and Privacy’ covers these issues in detail.

Collection of data
With a firm specification of the data, the collection should become a straightforward task for each participant. In a pilot project where everything is new, there are inevitable slips and problems, but the experience we have gained from finding solutions is available for future exercises and to future participants.

Extraction of data becomes a reproducible task, so a participant can re-run the same database query to generate a new report for the current stock and tenants over an appropriate period.

Participants using similar software systems to manage their stock can also collaborate in devising methods to most easily generate these reports. This is best managed through a working group approach.

Integration and validation of data
Integration of data was achieved by loading the participants’ reports into a combined database. At that stage, HouseMark performed some basic data quality and consistency checks, such as verifying the total number of properties, confirming that rents appear plausible (which they might not if a participant has, for example, accidentally reported monthly rents instead of weekly rents).

External data
A landlord’s data systems will naturally store only information about the landlord’s properties and tenants. The power of Big Data often comes from integrating data from different sources; these are sometimes seen on web sites as ‘mash-ups’. Constructing these linkages is often a complex task, best performed by specialists, such as HouseMark and ADAC.

In the pilot project, some of the linkages to postcode data such as deprivation statistics were relatively crude. With extra effort, it will be possible to improve the spatial resolution of some data, leading to better analyses.
Analysis
The initial analysis phase usually consists of visualising and understanding the data set. Cluster analysis is an ideal form of exploratory data analysis, but it is not the only one. The best form of analysis depends on the questions being asked. Is the aim simply to identify patterns in the data? Is the aim to find a means of predicting some variable? Is the aim to monitor changes from one period to the next?

Time spent simply looking at the data is rarely wasted. Unexpected patterns might indicate problems with the data collection or integration, or they might suggest new insights or questions to be explored.

Interpretation
There is no purpose in manipulating and visualising data unless it leads to useful knowledge and ultimately to insights that can affect how an organisation acts.

Without specialist knowledge, a data analyst can easily be misled by strong correlations that are not actually useful. For example, if an area has a low level of rent arrears, it will also have a low level of court actions for recovery of arrears. So that correlation is not a useful insight because both reflect a common confounding (but hidden) variable: difficulty in paying rent.

HouseMark’s analysts have the specialist housing knowledge which allows them to untangle these issues and look for the useful and interesting insights.

Bonus side-effects
With modern computers and software systems, handling large amounts of data is straightforward. There is usually no penalty for collecting extra data items. This means that a data set built to answer one question can often be wide enough and deep enough to be explored for other topics. For example, this data set was investigated with regard to repairs management, but it can also be used to investigate rent arrears issues, or occupancy issues.

Low cost experiments
One use of Big Data analytics that comes from knowing that certain features exist within the data is the possibility of designing experimental interventions. If we know that there are certain high risk categories, then they can be given priority in testing methods, e.g. support at the start of a tenancy and frequent follow-ups.

This optimises use of resources – but a key benefit that collaboration with other HouseMark members brings is the ability to aggregate results across organisations to see what works and what does not, i.e. use a membership-scale sample not a member-scale sample.

Discussion

There is no purpose in manipulating and visualising data unless it leads to useful knowledge and ultimately to insights that can affect how an organisation acts.
Explore your own data

How HouseMark can help

HouseMark can help you apply the Big Data techniques contained in this report to your own datasets, or datasets shared between yourself and other landlords.

Using HouseMark’s consultants and analysts, alongside our data scientist associates from the Advanced Data Analytics Centre of Nottingham University, and our legal partners Anthony Collins, we can help you with:

• extracting and preparing your data for analysis, achieving consistency in definitions if pooling data from a group of landlords
• identifying relevant datasets from open data and other sources to complement your own datasets in the analysis
• managing data protection and anonymisation issues appropriately and within the law
• identifying the appropriate analytical techniques for your data, and the knowledge you want to gain from it
• conducting the analysis for you from a full range of innovative techniques, and working with you to interpret the findings and make sense of them in your ‘real world’ situation
• identifying funding opportunities for Big Data innovation that would be relevant to your project.

To find out more, please get in touch with HouseMark:
T: 024 7647 2703
E: consultancy@housemark.co.uk.
### Appendix A

**Property locations - list of local authorities**

- Ashford
- Babergh
- Barnet
- Bedford
- Bexley
- Blaby
- Braintree
- Breckland
- Brighton and Hove
- Canterbury
- Corby
- Coventry
- Croydon
- East Northamptonshire
- East Staffordshire
- Eastbourne
- Epping Forest
- Epsom and Ewell
- Great Yarmouth
- Greenwich
- Hammersmith and Fulham
- Hastings
- Hinckley and Bosworth
- Ipswich
- Kensington and Chelsea
- Kettering
- Lewes
- Lichfield
- Maidstone
- Medway
- Milton Keynes
- Newcastle-under-Lyme
- North Norfolk
- North Warwickshire
- Northampton
- Norwich
- Nuneaton and Bedworth
- Reigate and Banstead
- Rother
- Rugby
- Sevenoaks
- Shepway
- South Norfolk
- Stoke-on-Trent
- Suffolk Coastal
- Sutton
- Thanet
- Tonbridge and Malling
- Tunbridge Wells
- Waveney
- Waverley
- Wellingborough
Appendix B
List of landlord data collected

Appendix C
List of data used in clustering

---

**Stage 1 variables**
- Property reference
- Postcode sector
- Stock type
- Whether self-contained or a bedspace
- Number of bedrooms
- Void stats
- Weekly net rent
- Weekly service charge eligible for HB
- Weekly service charge not eligible for HB
- Total weekly charge (calculated)
- Date of last re-let
- Number of repairs in 2013/14
- Number of terminations in 2013/14
- Number of days void in 2013/14
- Cost of responsive repairs in 2013/14

**Stage 2 variables**
- Dwelling type
- Form of structure
- Main heating type
- Floor level
- Year built
- Type of construction
- Existing use value for social housing
- Current SAP rating
- Whether or not charged as security against loan portfolio
- Net present value
- Whether purpose-built or converted
- Number of bedspaces
- Floor area
- Whether or not property has a garage
- Current environmental impact rating
- Cyclical repairs cost in 2013/14
- Capitalised major works expenditure in 2013/14

**Stage 3 variables: gathered at postcode sector level**
- Whether or not serious and less serious ASB has occurred
- Number of properties using specific rent collection methods
- Number of properties in specific rent arrears bands
- Number of properties at specific levels of rent arrears action
- Number of properties with known Universal Credit claimants
- Number of properties with tenants affected by the bedroom tax
- Number of properties with tenants affected by the benefits cap
- Number of properties where lead tenant ethnicity is known to be White British or not White British

**Landlord data**
To ensure comparability of landlord-supplied data, the raw figures for all of these measures was divided by the total properties in the dataset from that postcode sector and expressed as a percentage of total properties in the dataset from that postcode sector.

**Stage 3 variables: gathered at postcode sector level**
- Whether or not serious and less serious ASB has occurred
- Number of properties using specific rent collection methods
- Number of properties in specific rent arrears bands
- Number of properties at specific levels of rent arrears action
- Number of properties with known Universal Credit claimants
- Number of properties with tenants affected by the bedroom tax
- Number of properties with tenants affected by the benefits cap
- Number of properties where lead tenant ethnicity is known to be White British or not White British

**Percentage of properties in postcode sector with each of the following welfare benefit statuses at year end:**
- properties with Universal Credit claimants
- no known Universal Credit claimants
- tenants affected by Bedroom Tax
- no known tenants affected by Bedroom Tax
- tenants affected by the benefits cap
- no known tenants affected by the benefits cap

**Percentage of properties in postcode sector where the lead tenant at year end is known and has each of the following ethnicities:**
- White British
- not White British

**Percentage of properties in postcode sector where the lead tenant at year end is in each of the following bands:**
- 24 and under
- 25-34
- 35-44
- 45-54
- 55-64
- 65-74
- 75-84
- 85+

---

**Exploring clusters in social housing data**
- Percentage of properties in postcode sector with each of the following welfare benefit statuses at year end:
- properties with Universal Credit claimants
- no known Universal Credit claimants
- tenants affected by Bedroom Tax
- no known tenants affected by Bedroom Tax
- tenants affected by the benefits cap
- no known tenants affected by the benefits cap

**Percentage of properties in postcode sector where the lead tenant at year end is known and has each of the following ethnicities:**
- White British
- not White British

**Percentage of properties in postcode sector where the lead tenant at year end is in each of the following bands:**
- 24 and under
- 25-34
- 35-44
- 45-54
- 55-64
- 65-74
- 75-84
- 85+
Appendix C
List of data used in clustering

External data
Crime figures from Crime Survey for England and Wales:
- total crime
- homicide
- sexual offences
- violence with injury
- violence without injury
- possession of weapons offences
- robbery
- theft from the person
- bicycle theft
- shoplifting
- all other theft
- domestic burglary
- non-domestic burglary
- criminal damage and arson
- drug offences
- fraud offences
- public order offences
- vehicle offences
- miscellaneous crimes against society

Deprivation data from the English Indices of Deprivation 2010.

Appendix D
Acknowledgements

This discovery project has been informed by valued contributions from the sector. We acknowledge and thank the people involved.

In particular, we wish to thank Ross Fraser (formerly HouseMark Chief Executive) and Liz Kenny for editing this publication.

We also thank Paul Nicholson, Head of Business Performance at Hastoe Group, for acting as a ‘critical friend’ on drafts of the publication.

Project round tables
Landlords who contributed data to the discovery project helped steer the direction of the analysis at round tables. Attendees at these round tables were:
- Dean Ballard, Head of Performance Excellence, Orbit
- Sharon Palfreyman, Performance Consultant, Orbit
- Trish Thomas, Information Governance Specialist, Orbit
- Steve Rice, Head of Housing, DCH Group
- Ying Shao, formerly Information Analyst, Aspire
- Phil Scott, formerly Performance Manager, mhs homes.