infas 360

infas 360 GmbH

Ollenhauerstraße 1
D-53113 Bonn
Tel: +49 (0)228/74887-360
Fax: +49 (0)228/74887-399
E-Mail: info@infas360.de
www.infas360.de

Qualität von Geomarketing-Daten mit wissenschaftlichem Anspruch

White Paper, Stand März 2016

1. Einführung

Auch in der Wissenschaft wächst zunehmend der Bedarf, Befragungsdaten um zusätzliche mikrogeographische Merkmale anzureichern, um bessere, aussagefähigere und objektivere Ergebnisse zu erzielen. Die dazu bisher am Markt verfügbaren und angebotenen Daten wurden den wissenschaftlichen Qualitätsansprüchen allerdings oftmals nicht gerecht.

Die Qualität mikrogeographischer Daten definiert sich vor allem durch die nachfolgenden Parameter:

1. **Transparenz der verwendeten Rohdatenquelle(n)**
2. **Nachvollziehbarkeit und Reproduzierbarkeit der Berechnungsmethodik**
3. **Aussagekraft des Merkmals (Güte der Vorhersage)**
4. **Vollständigkeit des Datensatzes**
5. **Räumliche Ebene und Lagegenauigkeit**
6. **Aktualität und Aktualisierungszyklen**

In Zeiten von „Big Data“ entstehen fortlaufend neue Informationsquellen auf Basis der zunehmenden Anzahl verfügbarer amtlicher und privatwirtschaftlicher Datenbestände. Dieser Trend verstärkt sich durch den mittlerweile deutlich vereinfachten Zugang bzw. Zugriff auf amtliche Datenquellen.

Damit einher geht aber auch das Fehlen sogenannter „Metainformationen“ (Informationen über die Informationen), die Aufschluss über die o.g. Qualitätsparameter geben können. Als Beispiel soll hier der mikrogeographische Rasterdatensatz der amtlichen Einwohnerzahlen von 2011 auf Ebene von 100x100 Metern genannt sein. Positiv ist, dass im Rahmen der europäischen Open und Public Data- Bestrebungen überhaupt ein solcher Datensatz kostenlos zur Verfügung steht. Allerdings existieren über dessen Entstehen und Qualität kaum Angaben. Streng wissenschaftlich gesehen ist der Datensatz also nicht einsetzbar.

Ähnlich verhält es sich mit den am Markt angebotenen Daten aus privatwirtschaftlichen Quellen. Eine Vergleichsstudie der relevanten verfügbaren regionalen Kaukraftzahlen mit den Zahlen der amtlichen Einkommenssteuerstatistik hat beispielsweise ergeben, dass die auf Basis der Einkommenszahlen berechneten regionalen Prognosen einer am Einkommen gemessenen Kaufkraft Ergebnisse erbringen, die wenig belastbar sind.

Deshalb hat sich infas 360 in Abstimmung mit dem infas Institut für angewandte Sozialwissenschaften dazu entschieden, unter dem Gesichtspunkt der o. g. Qualitätsparameter nahezu alle mikrogeographischen Marktdaten neu zu berechnen, so dass sie wissenschaftlichen Ansprüchen genügen.

2. Die Neuberechnung der Mikrogeographie auf amtlicher Basis

Grundlegende Idee war die Schaffung einer flächendeckenden mikrogeographischen Einheit maximaler Granularität auf amtlicher Basis, die jedes Phänomen auf der Erdoberfläche beschreiben kann. Das Ergebnis ist eine amtliche Objektdatenbank, die alle amtlichen Gebäude und die sogenannten Points-Of-Interest (POIs) beinhaltet:

Die amtlichen Gebäude umfassen

- alle Wohnhäuser und Firmengebäude mit postalischen Adressen sowie
- alle weiteren Gebäude und Bauwerke wie z.B. Lagerhallen, Garagen oder Sportarenen, die aus dem Kataster stammen und nur teilweise über Adressen verfügen.

Die POIs beinhalten den Objekttyp bzw. die Nutzungsart dieser Gebäude und dienen vor allem Versorgungsanalysen wie z. B. das Finden der nächstgelegenen Schule, Polizeistation, Sehenswürdigkeit, Freizeitanlage oder dem nächsten Funkmast.

Alle amtlichen Objekte (mit oder ohne postalische Adresse) werden über eine Lage (Koordinate) definiert und fortlaufend in einer (Geo-)Objektdatenbank gepflegt und vervollständigt.

Wie Gefäße werden diejenigen Objekte, die postalische Adressen besitzen, fortlaufend mit Merkmalen wie folgt befüllt:

a. Quellen und Verfügbarkeit

Die Objektdatenbank wird gespeist aus den rund 50 Mio. 3D-Gebäuden der Bundesländer (HU und LOD1), der rund 22 Mio. Katasterkoordinaten mit postalischen Adressen der Länder (HK) sowie der ebenfalls rund 22 Mio. Anschlussobjekte der Lokationsdatenbank der Deutschen Telekom.

Die Füllung der Gefäße erfolgt mit Daten aus gesicherten Quellen wie z. B. das BKG, Destatis, die Statistischen Landesämter, Haushaltsinformationen der Deutschen Post AG, Firmendaten der Bisnode AG (ehemals Hoppenstedt) und Immobiliendaten von Immobilienscout24. Über das Data Intelligence Network stehen weitere Daten zur Verfügung.

Die Datenquellen zu jedem Merkmal werden grundsätzlich benannt. Auf Wunsch können neue Merkmale jederzeit über neue Datenquellen gerechnet werden.

b. Methodische Nachvollziehbarkeit und Reproduzierbarkeit

Über nachvollziehbare Berechnungsmethoden (Regeln) werden aus den Rohdaten neue Merkmale generiert. So wird z.B. auf Basis der rund 50 Mio. amtlichen 3D-Gebäudedaten mittels der Gebäudeparameter und einer Clusteranalyse ein statistisch hoch differenzierender Wohngebäudetyp entwickelt. Bei Bedarf und berechtigtem Interesse können jederzeit grundlegende Berechnungsregeln zu allen selbstentwickelten Merkmalen offen gelegt werden.

In diesem Kontext hat sich die infas 360 vor allem auf die Verfahren der Small-Area-Methoden (SAM) und das Duplizieren von Daten über die Bestimmung statistischer Zwillinge spezialisiert.

Small Area Methoden

Mit Small-Area-Methoden (SAM) können für jede geographische Ebene Mittelwerte, Anteile, Proportionen oder Raten geschätzt werden. Besonders interessant sind Small-Area-Methoden für die Regionalisierung amtlicher Daten wie z.B. Einkommen oder Arbeitslosenzahlen. Diese liegen meistens nur auf eher groben räumlichen Ebenen der Kreise oder Gemeinden vor. Mit Hilfe von Befragungsdaten (dabei sind nur geringe Fallzahlen erforderlich) lassen sich diese Werte auf kleinere Einheiten, wie den Ortsteil oder den Siedlungsblock rechnen. Das Schätzen von Marktanteilen auf kleinräumiger Ebene kann ebenfalls auf diese Weise verlässlich gerechnet werden.

Die Vorteile von SAM gegenüber herkömmlichen Schätzmodellen bestehen darin, dass in das Modell Informationen aus übergeordneten Raumebenen ebenso einfließen wie Informationen aus ähnlichen Regionen. Das führt zu sehr genauen Angaben, selbst wenn die Regionen nur kleine oder gar fehlende Fallzahlen aufweisen. Voraussetzung der Raumebenen ist eine hierarchisch überschneidungsfreie Struktur (‚nested data'). Dabei wird angenommen, dass bestimmte Zusammenhänge zwischen Variablen nicht nur in einem Gebiet Gültigkeit besitzen, sondern generell in der Population auf ähnliche Weise vorliegen. Somit können Informationen aller oder zumindest vieler Regionen zu einer Stabilisierung der Schätzungen in den einzelnen Regionen beitragen.

PAGS - das Postalisch-Amtliche Gliederungssystem

Im Zusammenhang mit den Anforderungen der ‚nested areas' hat infas 360 erstmals ein durchgängiges Postalisch-Amtliches Gliederungssystem auf (fast ausschließlich) amtlicher Datenbasis erstellt. Dazu zählen die bereits unter Punkt 2 erwähnten 50 Mio. 3D-Gebäudedaten in Verbindung mit 22 Mio. Katasterkoordinaten mit postalischen Adressen als kleinste Objekteinheit. Damit verschmolzen wurden bundesweit alle ca. 3,5 Mio. amtlichen Siedlungsblöcke, die gleichzeitig zur Generierung von ca. 75.000 Stadtvierteln und Ortsteilen unterhalb der Gemeindeebene (=AGS8 Amtlicher Gemeindeschlüssel, achtstellig) dienen.

Die resultierenden räumlichen Ebenen durch das postalisch-amtliche Gliederungssystem in Deutschlands sehen wie folgt aus

AGS8 = Amtlicher Gemeindeschlüssel
AGS11 = Stadt- und Ortsteilschlüssel
AGS16 = AGS11 + 5-stellige Postleitzahl
AGS20 = Amtlicher Siedlungsblock (AGS11 + 7 stellige amtliche ID)
AGS22 = Amtliche Straßenblockseite (AGS20 + Straßenabschnitts-ID)
AGS25 = Amtlicher Gebäudepunkt des amtlichen Gebäudes
AGS27 = Postalisch-amtliche Adresse in einem Gebäude

Je nach Datenquelle und Merkmalstyp liegen die mikrogeographischen Daten auf einer oder mehrerer dieser räumlichen Ebenen vor (z.B. die Anzahl der Einwohner AGS8 bis AGS27).

Zwillingssuche und Datenduplizierung

infas 360 nutzt zur Übertragung von Daten den sogenannten „Similarity Score“ sowie das Nachbarschaftsverhältnis von Objekten. Der Similarity Score drückt die Ähnlichkeit von Gebäudeobjekten aufgrund ihrer vorhandenen Gebäudeparametern (z.B. Komplexität des Grundrisses und Gebäudehöhe/-volumen) untereinander aus.

Des Weiteren wird die räumliche Nähe von Objekten zueinander bewertet. Dafür werden je nach Anforderung i.d.R. zwei Verfahren eingesetzt:

1. Ein Wert wird dupliziert innerhalb eines amtlichen Siedlungsblock (AGS20 mit ca. 20 Haushalten im Mittel)

2. Ein Wert wird mittels räumlicher Interpolation und einer bestimmten Ausbreitungsmethode (grundsätzlich gilt: je weiter weg, desto weniger) in die Fläche übertragen.

c. Güte und Aussagekraft

Datenquelle und Berechnungsmethode führen zu einem Merkmal, dass schlussendlich nichts anderes darstellt als eine Schätzung eines bestimmten Phänomens auf einer bestimmten räumlichen Ebene (z.B. die Anzahl der Einwohner in einem Siedlungsblock, dem AGS20) zu einem bestimmten Zeitpunkt/Stichtag. Die Güte der Vorhersage ist das A und O der Datenqualität. Diese lässt sich entweder durch den Vergleich mit Echtdaten (Begehungs- und/oder Befragungsdaten) und/oder durch den Abgleich mit Drittquellen (durch Recherche) bestimmen.

Die Bestimmung der Datengüte wird fortlaufend und stichprobenartig durchgeführt und variiert von Merkmal zu Merkmal. Sie kann auf Nachfrage erstellt und zur Verfügung gestellt werden.

d. Vollständigkeit

Der Datensatz ist flächendeckend und hat den Anspruch stichtagsbezogen vollständig zu sein (d.h. zu einem Tag x befinden sich alle zu diesem Zeitpunkt bekannten Gebäuden aller Gemeinden Deutschlands in der Objektdatenbank).

Trotzdem können von 1.000 Befragten i.d.R. nicht 100% aller Adressen angereichert werden. Ein jeweiliges Geocodierungsprotokoll gibt Aufschluss über die Vollständigkeit eines Datensatzes. Vollautomatisch werden rund 95% aller Fälle angereichert. Je nach Adressqualität differiert der Wert zwischen 93% und 97%. Über manuell unterstützte Abgleichsverfahren kann die Vollständigkeit einer Datenanreicherung auf bis zu 99% der Fälle angehoben werden. Durch eine reine manuelle Nachcodierung kann projektweise auch eine 100%ige Vollständigkeit gewährleistet werden.

e. Aktualität

Alle Daten beziehen sich auf den jeweiligen Stand der jeweiligen Quelle sowie der für die Berechnung zu Grunde liegende PAGS-Gebietsstand. PAGS und die damit verbundenen geographischen Gebiete wie auch die dazugehörigen Merkmale beziehen sich auf die aktuell

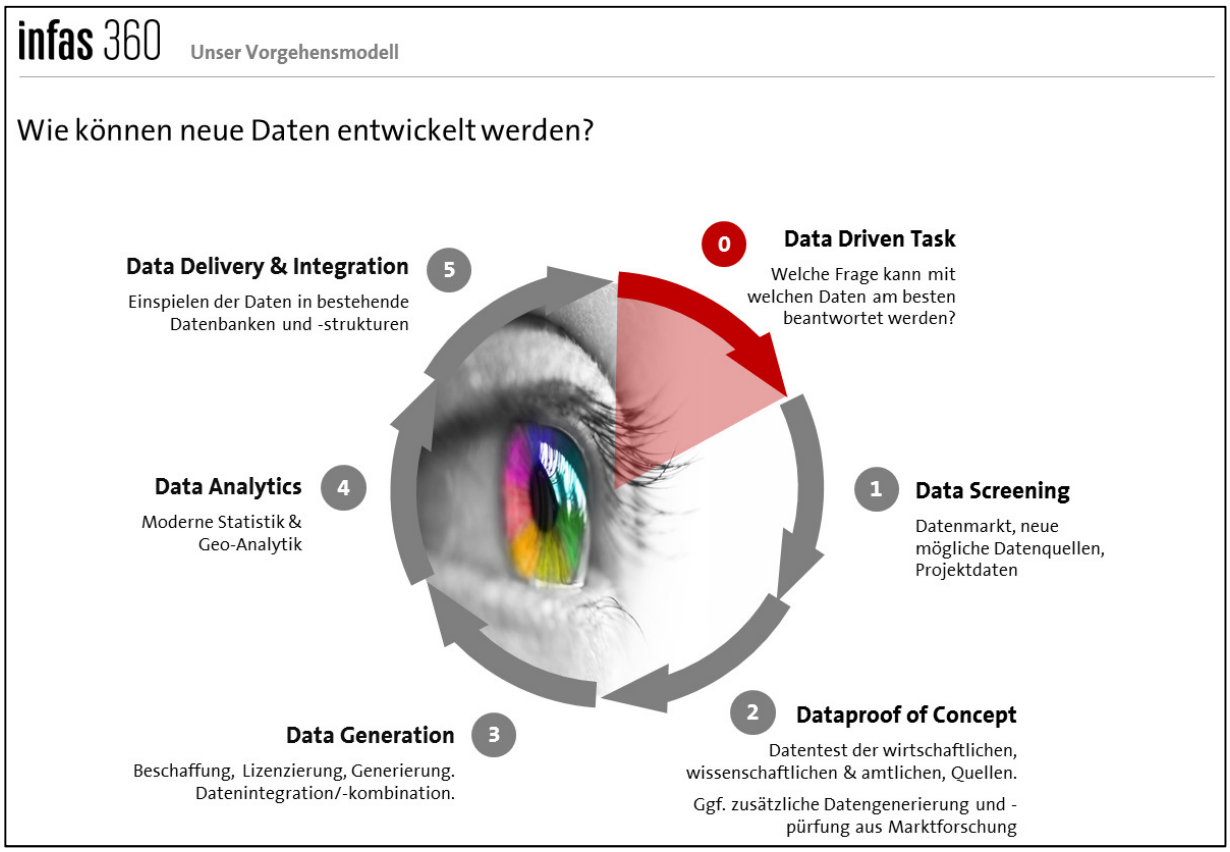
zur Verfügung stehende amtliche Statistik (= 31.12. des Vorvorjahres) und steht Ende des 1. Quartals eines Jahres zur Verfügung. Das bedeutet, dass 2016, die Daten den Stand der Statistik vom 31.12.2014 beinhalten. Dies ist dem Umstand geschuldet, dass die amtliche Statistik erst am Ende eines Folgejahres veröffentlicht wird. Die amtlichen Geometrien sind i.d.R. aktueller, beziehen sich jedoch aus Konsistenzgründen auf denselben Stichtag (hier 31.12.2014).

Die Daten werden (intern) fortlaufend gepflegt und aktualisiert. Projektbezogen können sowohl unterjährige als auch aktuellere Gebietsstände geliefert werden.

3. Projektbezogene Datenentwicklung

In Zeiten von Big Data wachsen die Datenquellen und die daraus verfügbaren Daten. Dies führt zu einem Trend, dass Fragestellungen nicht nur mit Hilfe eines allgemeinen Standarddaten-portfolios beantwortet werden, sondern zunehmend auch über individuelle projektbezogene Daten, die ad-hoc recherchiert und erstellt werden.

Nachfolgend aufgeführt wird ein idealisiertes Vorgehensmodell für eine fragespezifische Datenentwicklung, die über das Standardportfolio hinausreicht.



4. Der Standarddatenkatalog

Der Standarddatenkatalog beinhaltet alle Merkmale und Variablen, die mindestens einmal jährlich aktualisiert und zum Gebietsstand 31.12. des Vorvor-Jahres referenziert werden. Auf Wunsch können unterjährige Aktualisierungen erfolgen, gehören aber nicht zum Standard.

Hingegen können weitere, zusätzliche Merkmale und Variablen auch unterjährig hinzukommen, da fortlaufend neue Daten entwickelt werden. Die aktuellste Liste aller Merkmale aus dem Standarddatenkatalog befindet sich in der Datei

- infas360_Datenbeschreibung_CASA_Jahr_Monat_Tag

5. Kontakt & Ansprechpartner

Kaufmännisch
Tobias Gödderz
Head of Consulting Geomarketing
t.godderz@infas360.de
0049 – 228 74887 – 373

Inhaltlich
Dr. Barbara Wawrzyniak
Leiterin Daten & Analysen
b.wawrzyniak@infas360.de
0049 – 228 74887 – 369

infas 360 GmbH, Ollenhauer Str. 1, 53113 Bonn, www.infas360.de