

MOSAIC Milieus^{®1} - Übertragung eines psychographischen Zielgruppenmodells mittels multinomialer logistischer Regression in den Raum

Sascha Mertes
microm Micromarketing-
Systeme und Consult GmbH
Hellersbergstr. 11
41460 Neuss
s.mertes@microm-online.de

Zusammenfassung

Psychographische Zielgruppenmodelle berücksichtigen über sozioökonomische und soziodemographische Zusammenhänge hinaus Wertvorstellungen und Einstellungen der Konsumenten, die deren Konsumverhalten maßgeblich beeinflussen. Die MOSAIC Milieus[®] sind eine Adaption der Sinus-Milieus[®] auf microgeographischer Ebene und bieten dem Anwender die Möglichkeit, diesen Aspekt für wirkungsvollere Marketingaktionen zu nutzen. Der folgende Artikel wirft einen Blick hinter die Kulissen der microgeographischen Datenerstellung und zeigt, wie die Modellierung der MOSAIC Milieus[®] erfolgt ist und welche Aspekte bei der Auswahl geeigneter Verfahren und Variablen berücksichtigt werden mussten.

Schlüsselwörter: Sinus Sociovision, Sinus-Milieus[®], microm, MOSAIC Milieus[®], Mikrogeographie, MOSAIC Datensystem, Multinomiale logistische Regression, Variablenauswahl, Collapsing levels, Variable Clustering

1 Einführung in die Sinus-Milieus[®]

Die Sinus-Milieus[®] sind das Ergebnis von fast 30 Jahren sozialwissenschaftlicher Forschung. Die Zielgruppenbestimmung von Sinus Sociovision orientiert sich an der Lebensweltanalyse unserer Gesellschaft. Die Sinus-Milieus[®] gruppieren Menschen, die sich in ihrer Lebensauffassung und Lebensweise ähneln.

Grundlegende Wertorientierungen gehen dabei ebenso in die Analyse ein wie Alltagseinstellungen zur Arbeit, zur Familie, zur Freizeit, zu Geld und Konsum. Sie rücken also den Menschen und das gesamte Bezugssystem seiner Lebenswelt ganzheitlich ins Blickfeld. Sie bieten deshalb dem Marketing mehr Informationen und bessere Entscheidungshilfen als herkömmliche Zielgruppenansätze.

¹ © MOSAIC Milieus[®] ist ein Gemeinschaftsprodukt der Sinus Sociovision GmbH, Heidelberg und microm Micromarketing-Systeme und Consult GmbH, Neuss

Die Sinus-Milieus[®] werden seit Beginn der 80er Jahre von führenden Markenartikel-Herstellern und Dienstleistungsunternehmen für das strategische Marketing, für Produktentwicklung und Kommunikation erfolgreich genutzt. Große Medienunternehmen arbeiten damit seit Jahren genauso wie Werbe- und Mediaagenturen.

Die Grenzen zwischen den Milieus sind fließend. Lebenswelten sind nicht so exakt eingrenzbar wie soziale Schichten. Dies wird auch als Unschärferelation der Alltagswirklichkeit bezeichnet. Ein grundlegender Bestandteil des Milieu-Konzepts ist, dass es zwischen den Milieus Berührungspunkte und Übergänge gibt. Diese Überlappungspotentiale sowie die Position der Milieus in der Gesellschaft nach sozialer Lage und Grundorientierung veranschaulicht die folgende Grafik: Je höher ein Milieu in dieser Grafik angesiedelt ist, desto gehobener sind Bildung, Einkommen und Berufsgruppe; je weiter rechts es positioniert ist, desto moderner ist die Grundorientierung. In dieser "strategischen Landkarte" können Produkte, Marken, Medien etc. positioniert werden.

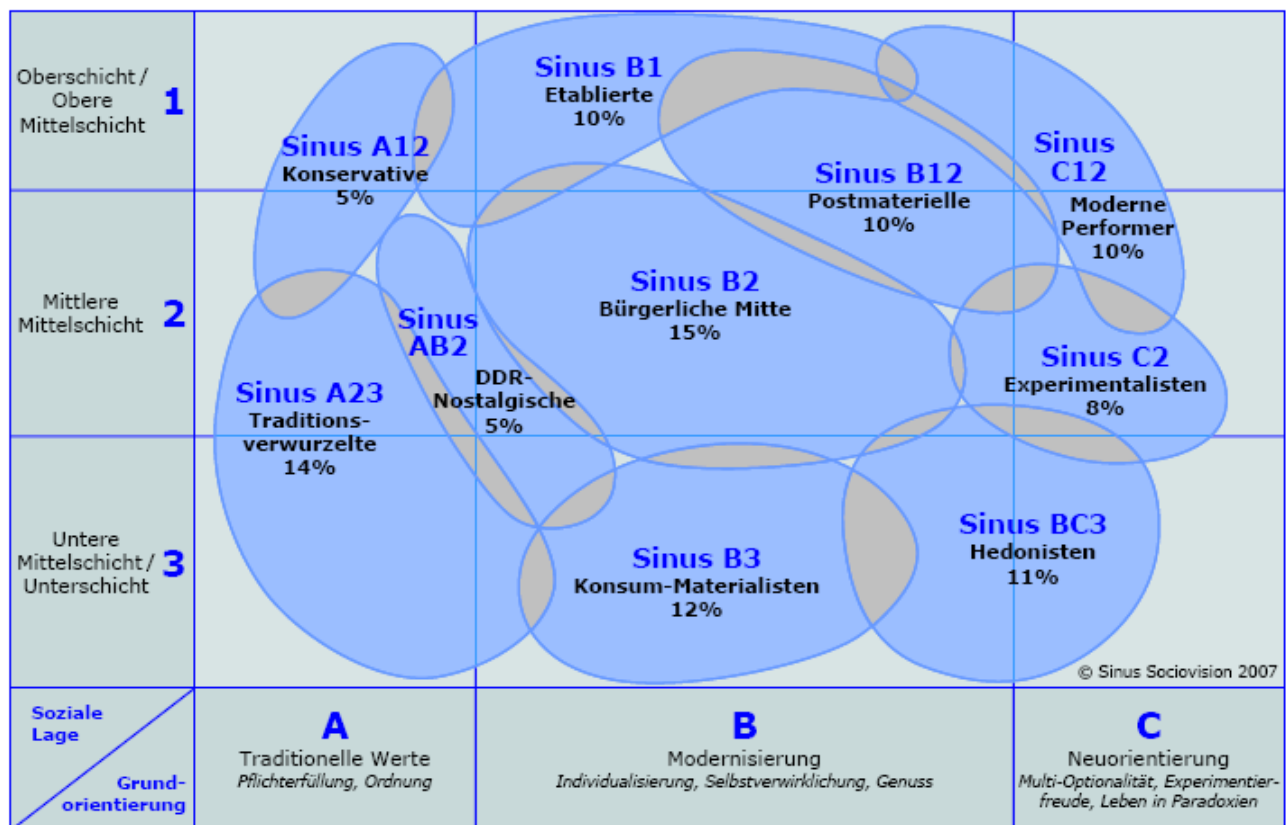


Abbildung 1: Die Sinus-Milieus[®] in Deutschland – Soziale Lage und Grundorientierung

2 Die Idee der MOSAIC Milieus[®]

Um die Sinus-Milieus[®] für räumliche Planung und Marketing-Aktivitäten nutzbar zu machen, haben Sinus Sociovision und microm gemeinsam die MOSAIC Milieus[®] entwickelt. Zu diesem Zweck wurden im Jahr 1998 erste Tests durchgeführt, um Zusammenhänge zwischen der MOSAIC Datenbank und den Sinus-Milieus[®] zu überprüfen.

Nachdem diese erfolgreich verlaufen waren, gab es 1999 die ersten MOSAIC Milieus[®] für das westdeutsche Milieu-Modell von Sinus Sociovision. Dadurch erhielten die Sinus-Milieus[®] über ihre strategische Aussage hinaus erstmalig auch einen lokalisierbaren Nutzen.

Nachdem Unternehmen wie die Deutsche Bank, TUI und Tchibo erfolgreich mit den MOSAIC Milieus[®] gearbeitet hatten, entschied man sich, auch das im Oktober 2001 veröffentlichte gesamtdeutsche Milieu-Modell in den Raum zu übertragen. Die MOSAIC Milieus[®] sind seitdem für ganz Deutschland verfügbar und werden regelmäßig, das heißt in einem jährlichen Zyklus, aktualisiert.

3 Gleich und gleich gesellt sich gern: Das MOSAIC Datensystem

Zwischen der räumlichen und der sozialen Nähe von Menschen gibt es deutliche Zusammenhänge. Menschen, die im gleichen Umfeld wohnen gehören meist auch einer gleichartigen sozialen Schicht an und weisen ähnliche Bedürfnisse und Konsumgewohnheiten auf.

Die Qualität einer microgeographischen Marktsegmentierung hängt von zwei Elementen ab: von der Erklärungskraft der inhaltlichen Merkmale und von der Feinheit des räumlichen Rasters. In MOSAIC sind Daten vor allem zu den folgenden Sachverhalten eingegangen, die entweder den Haushalt oder das weitere Umfeld beschreiben:

- Geographische Informationen, zum Beispiel die Lage im Raum bzw. im Stadtgebiet sowie Informationen zur Bebauungsstruktur (Haustyp, Straßentyp) etc.
- Ökonomische Informationen, zum Beispiel der soziale Status, die berufliche Qualifikation oder die statistische Wahrscheinlichkeit von Zahlungsausfällen etc.
- Demographische Informationen, zum Beispiel das Alter des Haushaltsvorstandes, die Wahrscheinlichkeit von Singlehaushalten etc.

Die Basisdatenbank ist ein Produkt der Kooperation mehrerer Partner, die ihre Datenbestände in anonymisierter Form eingebracht haben, wie zum Beispiel der Verband der Vereine Creditreform. Daneben werden auch amtliche Daten wie zum Beispiel PKW-Bestandsdaten des Kraftfahrtbundesamtes und Daten der Statistischen Landesämter aufbereitet und verarbeitet.

Diese Informationen liegen grundsätzlich für sämtliche (rund 40 Mio.) Haushalte Deutschlands vor und werden für die rund 17,8 Mio. Häuser ausgewiesen. Für die Analyse werden – aus Gründen des Datenschutzes – mehrere zu einem Wohnumfeld gehörende Häuser zu einem „virtuellen“ microgeographischen Segment gebündelt, das mindestens 5 und durchschnittlich 8 Haushalte umfasst. Dabei erfolgt eine „Begehung der Straßen im Computer“. Zunächst werden die Häuser danach geprüft, ob wenigstens 5 Haushalte darin vorkommen. Große Häuser bilden also eigene Segmente. Wo dies

nicht gegeben ist, werden in jeder Straße strukturähnliche Häuser zu Segmenten zusammengefasst, in denen zusammen mindestens 5 Haushalte vorkommen. Gleichzeitig werden die Einzeldaten zu aussagefähigen Strukturindikatoren für die Segmente verdichtet und die Information wieder pro Haus als Datei aufbereitet. Insgesamt fließen weit über eine Milliarde Einzelinformationen ein, die als MOSAIC Variablen aufbereitet für verschiedene Raumebenen entsprechend der folgenden Abbildung zur Verfügung stehen.

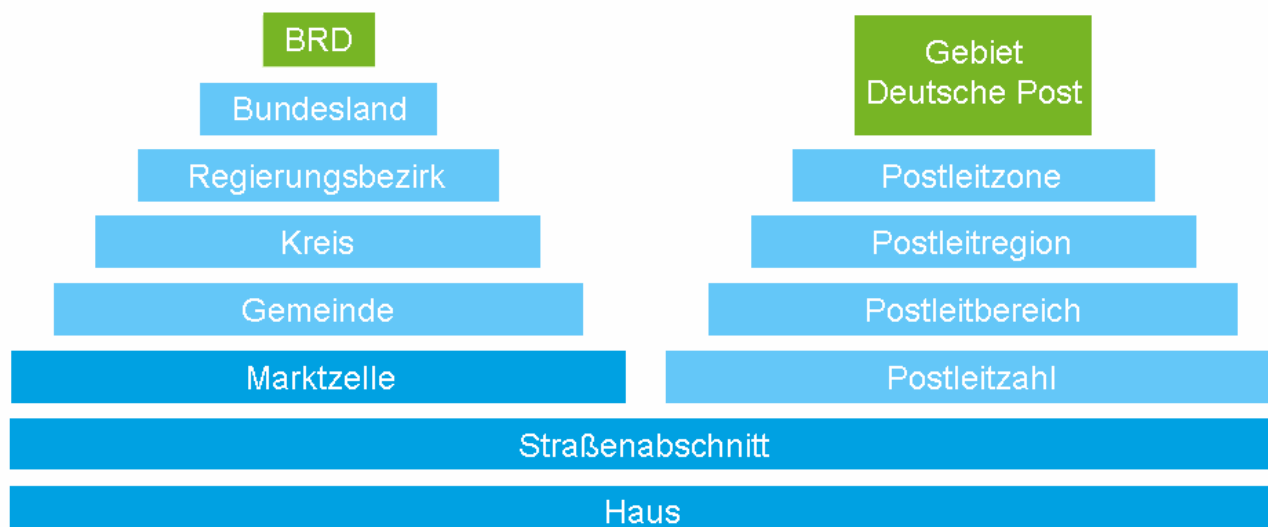


Abbildung 2: Übersicht der geographischen Gliederung und Verfügbarkeit der MOSAIC Variablen in Deutschland

4 Methodische Aspekte zur Variablenauswahl und zu Modellierungsstrategien

Für die Übertragung der Sinus-Milieus[®] in den Raum steht generell eine Vielzahl statistischer Methoden zur Verfügung. Die entscheidende Frage ist, welche Modellierungsstrategie die qualitativ besten Ergebnisse erwarten lässt. Ziel ist es, für jedes Haus in Deutschland einen Wert zu berechnen, der die Wahrscheinlichkeit des Auftretens eines spezifischen Sinus-Milieus[®] in diesem Haus angibt. Daneben soll ein sog. Dominantes Milieu pro Haus ausgewiesen werden, welches das besonders charakteristische Milieu eines Hauses repräsentiert. Rein statistisch betrachtet liegt hier also eine nominal skalierte, polytome abhängige Variable vor (Sinus-Milieus[®]), die mit unabhängigen Variablen unterschiedlichster Skalenniveaus (unter anderem MOSAIC Daten) vorhergesagt werden soll.

In Abhängigkeit von den technologischen Rahmenbedingungen sind im konkreten Fall folgende Verfahren zur Modellierung der Sinus-Milieus[®] in Betracht gezogen worden:

- Diskriminanzanalyse (DA)
- Multinomiale logistische Regression (MLR)
- Binäre logistische Regression (BLR)

Dabei gilt es zu beachten, dass bei der DA und der MLR keine Dichotomisierung der abhängigen Variable nötig ist, wohingegen die BLR per Definition eine dichotomisierte Zielvariable verlangt. Hier ist die Idee, jedes der zehn Sinus-Milieus[®] zunächst separat zu modellieren und anschließend in einem Gesamtmodell zu fusionieren. Man kann in diesem Fall auch von einem zweistufigen Modellansatz sprechen.

Grundsätzlich bieten sich alle drei Verfahren zur Modellierung der Sinus-Milieus[®] an, doch gibt es klare Vor- und Nachteile einzelner Methoden, die die Auswahl für eine der Methoden erleichtern. Vorteile der DA und der MLR sind unter anderem, dass es sich hier um Verfahren handelt, die mit einem überschaubaren Rechen- und Personalaufwand und somit in angemessener Zeit durchführbar sind, da sie alle zehn Sinus-Milieus[®] mit einem Modell abbilden können. Ein positiver Aspekt, der bei der BLR zum Tragen kommt, ist die Tatsache, dass jedes Sinus-Milieu[®] mit einer individuellen Kombination unabhängiger Variablen prognostiziert werden kann. Die Nachteile der Verfahren ergeben sich aus den jeweiligen Vorteilen. Die Prognose der Sinus-Milieus[®] mittels BLR erhöht den Zeitaufwand in einem nicht unerheblichen Maße und bei der DA und MLR werden alle zehn Milieus mit einer fixen Auswahl unabhängiger Variablen vorhergesagt, so dass hier ggf. auf ein Stück erklärter Varianz verzichtet werden muss. Ein weiterer Nachteil der DA ist, dass sie ausschließlich intervallskalierte Variablen verlangt. Im vorliegenden Fall wurde schließlich die MLR als Methode ausgewählt.

Neben der Wahl eines geeigneten statistischen Verfahrens zur Vorhersage der Zielgröße ist es natürlich auch notwendig, unabhängige Variablen bereitzustellen, die ein hohes Erklärungspotential besitzen. Dabei ist es oftmals so, dass eine schier unüberschaubare Zahl an unabhängigen Variablen zur Verfügung steht, die sich aus statistischer Sicht durch heterogene Qualität auszeichnen. Generell wird man hier sehr häufig mit folgenden Schwierigkeiten konfrontiert:

- große Anzahl an unabhängigen Variablen
- Missing Values
- Schiefe Verteilungen
- Variablen, die gleiche inhaltliche Dimensionen beschreiben
- Kategoriale Variablen mit sehr vielen Ausprägungen (hier bis zu 38 Kategorien)
- kein homogenes Skalenniveau bei den unabhängigen Variablen

Um Lösungsstrategien für die oben genannten Probleme zu entwickeln und schließlich die Frage zu klären, welche der unabhängigen Variablen für die Modellentwicklung berücksichtigt werden können und sollen, ist es unabdingbar sich mit folgenden Themen auseinander zu setzen. Dazu zählen unter anderem die Imputation, die Transformation und die Dimensionsreduktion. Aspekte, die dabei eine wichtige Rolle spielen, sind bei der Imputation der Ersatz fehlender Werte (missing values) mittels geeigneter Reprä-

sentanten und Verfahren wie zum Beispiel Mittelwert, Median, Regression Trees, verteilungsbasierte Methoden bei metrischen Variablen bzw. Modus, Konstanten, Entscheidungsbäume etc. für alle ordinal und nominal skalierten Variablen.

Die Transformation dient vor allem zur Optimierung der empirischen Verteilung und zur Varianzstabilisierung und Herstellung von Linearität und Normalverteilung mit Hilfe von Quadrieren, Logarithmieren, Kehrwerten oder Standardisierungen. Daneben zählt auch das Zusammenfassen von Ausprägungen aufgrund einer nicht optimalen Häufigkeitsverteilung (zum Beispiel rare events) dazu. Hier können auch Entscheidungsbäume ein geeignetes Verfahren darstellen.

Schließlich gilt es im Vorfeld Variablen zu erkennen, die gleiche oder ähnliche Dimensionen abbilden. Dabei helfen neben der einfachen Korrelationsanalyse alle strukturfindenden Verfahren wie beispielsweise die Faktoren- und Hauptkomponentenanalyse (PCA), Clusteranalyse, Self-Organizing Maps, Memory-Based Reasoning oder auch das Variable Clustering.

Letzteres soll neben den Entscheidungsbäumen zur Reduktion der Anzahl an Ausprägungen im Folgenden kurz näher betrachtet werden.

4.1 Variable Clustering als Methode zur Dimensionsreduktion

Wie bereits erwähnt ist es neben einer allgemeinen Reduktion der oftmals hohen Zahl an unabhängigen Variablen nötig, jene Variablen zu identifizieren, die eine starke Abhängigkeit untereinander aufweisen und somit gleiche oder ähnliche Dimensionen abbilden. Verfahren, die dieses leisten können, wurden weiter oben bereits genannt. Diese besitzen jedoch zum Teil deutliche Nachteile. So wird zum Beispiel bei einer einfachen Korrelationsanalyse nur die bivariate Abhängigkeit gemessen. Eine Faktoren- bzw. Hauptkomponentenanalyse bietet sich hier zwar generell an, ist jedoch aufgrund der neu erzeugten synthetischen Variablen im Nachhinein vielfach schwer interpretierbar. Ein Lösungsansatz stellt das Variablen-Clustering dar, welches SAS[®] mit der Prozedur `PROC VARCLUS` zur Verfügung stellt. Dieses Verfahren basiert auf Principal Components, also Hauptkomponenten und somit letztlich auf Korrelationen und findet Gruppen von Variablen (Cluster), die – entsprechend des allgemeinen Ziels eines Clusterings – mit ihrem Cluster stark und möglichst wenig mit Variablen aus anderen Clustern korrelieren. Die Strategie besteht nun darin, auf Basis von fachlichen Erwägungen jeweils eine Variable als Repräsentant aus dem Cluster auszuwählen. Der Vorteil gegenüber anderen Verfahren ist, dass hier multivariate Abhängigkeiten berücksichtigt werden und durch die Auswahl von „echten“ Variablen die Interpretierbarkeit des Modells auch später gewährleistet bleibt. Das Variablen-Clustering ist lt. Aussage von SAS[®] im neuen SAS[®] Enterprise Miner 5.3 standardmäßig implementiert.

4.2 Zusammenfassung von Ausprägungen mittels Entscheidungsbäumen

Die generelle Schwierigkeit bei unabhängigen Variablen mit vielen Ausprägungen bzw. Kategorien ist, dass selbst in größeren Stichproben schnell Fallzahlprobleme auftreten und die Dimensionalität im Modell extrem ansteigt. Ein möglicher Lösungsansatz ist

das Zusammenfassen einzelner Ausprägungen nominaler Inputs unter Berücksichtigung des Einflusses auf die Zielvariable mit Hilfe eines Entscheidungsbaumes, das sogenannte „collapsing levels“. Dabei ist zu beachten, dass bei mehreren nominal skalierten Variablen alle Inputs simultan kollabiert werden sollten, um Interaktionen und Abhängigkeiten (multivariate Beziehungen zwischen den unabhängigen Variablen und der Zielvariable) untereinander zu berücksichtigen. Die „Blätter“ dieses Baumes können im SAS[®] Enterprise Miner gespeichert und als Inputs für nachfolgende Verfahren verwendet werden.

Folgenden Einstellungen für den Entscheidungsbaum können empfohlen werden:

- Anzahl Branches entsprechend der Anzahl der Ausprägungen der Inputvariable
- maximale Tiefe des Baumes: 1 Ebene
- SAS[®]-Einstellungen: hoher p-Wert
- Signifikanzlevel= 1
- keine Bonferroni-Korrektur

5 Ein Sinus-Milieu[®] für jedes Haus: Das Vorgehen zur Berechnung der MOSAIC Milieus[®]

Für die Übertragung der Sinus-Milieus[®] in den Raum wurden von diversen Marktforschungsinstituten in Summe 35.000 Interviews durchgeführt um darüber eine Milieuzuordnung der Befragten vorzunehmen. Aus dieser Befragung konnten 33.000 Interviews verwendet werden, die eine eindeutige Milieuzuordnung aufwiesen und bei denen eine Referenzierung zum MOSAIC Datensystem möglich war. Die Referenzierung der Interviews in einer Datenbank erfolgte in mehreren Schritten unter Einhaltung der Datenschutzbestimmungen. Das Ergebnis ist eine anonymisierte Analysedatenbank auf Haus-ebene, in der die Milieuzugehörigkeit, also das Sinus-Milieu[®], als Zielvariable enthalten ist und die MOSAIC Daten als erklärende Variablen dienen.

Mittels der MOSAIC Daten, die Aussagen über den sozialen Status, Familienstruktur, Alter, Haustyp, Wohnumfeld, PKW-Nutzung, Konsumverhalten und Kaufkraft zulassen, wurden Zusammenhänge zur Milieuzugehörigkeit überprüft. Daneben spielten auch grobräumigere Daten wie Siedlungstypen und Wahlverhalten eine Rolle. So können zum Beispiel die Konservativen als statushohe, kaufkraftstarke, eher ältere Menschen beschrieben werden, die häufig Familien mit Kindern haben, ländliche oder Stadtrandlagen bevorzugen und die in ihrem Wahlverhalten eher zu konservativen Parteien neigen. Diese inhaltlichen Zusammenhänge lassen sich im Rahmen der Stichprobe von 33.000 Datensätzen mit Hilfe der multinomialen logistischen Regression messen und so auf andere Datenbestände übertragen.

Die folgende Abbildung fasst den Analyseprozess zusammen:

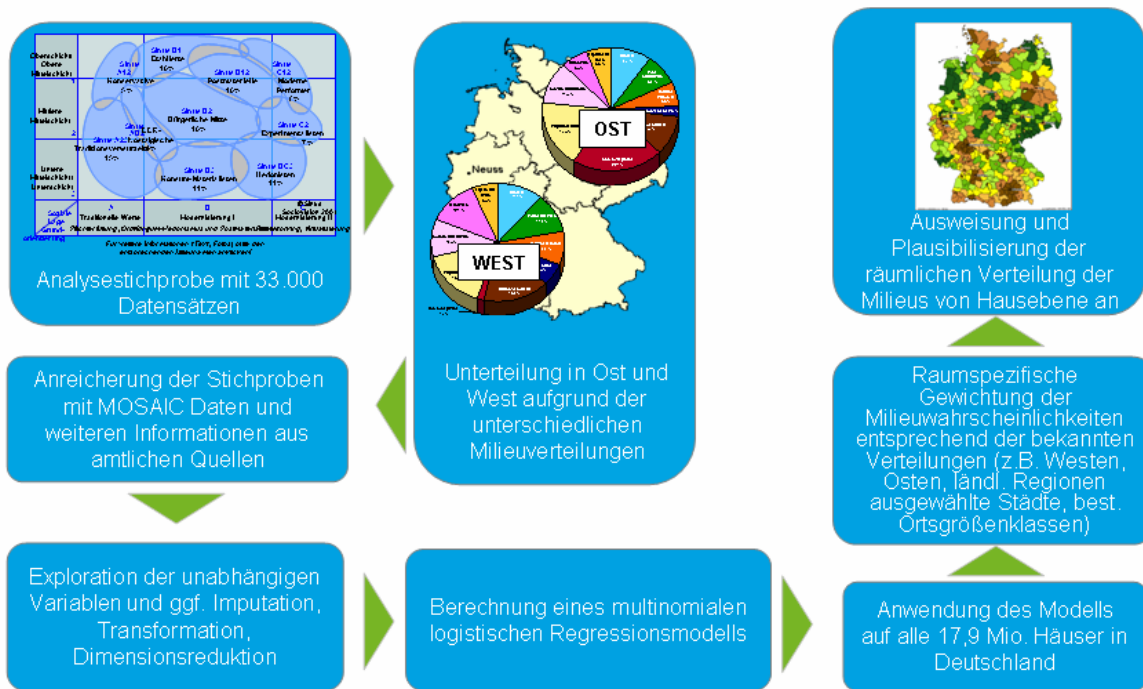


Abbildung 3: Der Analyseprozess

6 Das Ergebnis: Die MOSAIC Milieus®

Das Ergebnis dieses Modellierungsprozesses sind die MOSAIC Milieus®. Sie geben für jedes Haus in Deutschland an, mit welcher Wahrscheinlichkeit ein spezifisches Sinus-Milieu® dort vorzufinden ist. Darüber hinaus wurde ebenfalls ein sog. Dominantes MOSAIC Milieu® berechnet. Dieses gibt an, welches Sinus-Milieu® in einem Haus am stärksten überrepräsentiert ist und somit dieses besonders gut wiedergibt.

Durch die MOSAIC Milieus® ist es nun möglich, das Zielgruppenmodell der Sinus-Milieus® sowohl auf beliebige Kundenbestände als auch auf verschiedene Raumebenen – von Haus- über Straßenebene bis hin zu administrativen und postalischen Gliederungen – zu projizieren und so für Direktmarketingmaßnahmen nutzbar zu machen. Die Sinus-Milieus® erhalten so über ihre strategische Aussage hinaus einen lokalisierbaren Nutzen. Die folgenden Abbildungen 4-6 zeigen abschließend, wie sich die MOSAIC Milieus® auf unterschiedlichen Raumebenen unter anderem am Beispiel Heidelberg verteilen.

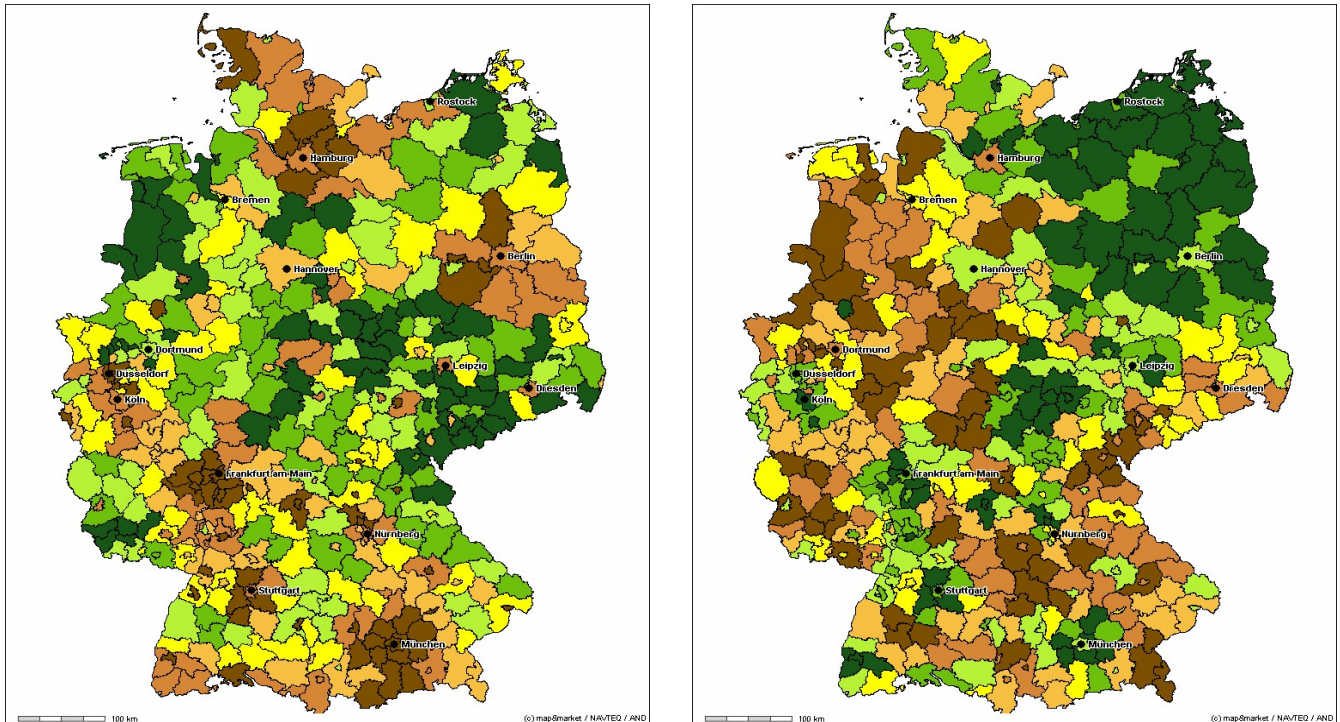


Abbildung 4: Wahrscheinlichkeiten der MOSAIC Milieus[®] auf Kreisebene für Moderne Performer (links) und Traditionsverwurzelte (rechts). Je brauner der Kreis, desto höher ist die Wahrscheinlichkeit das jeweilige Milieu dort vorzufinden.

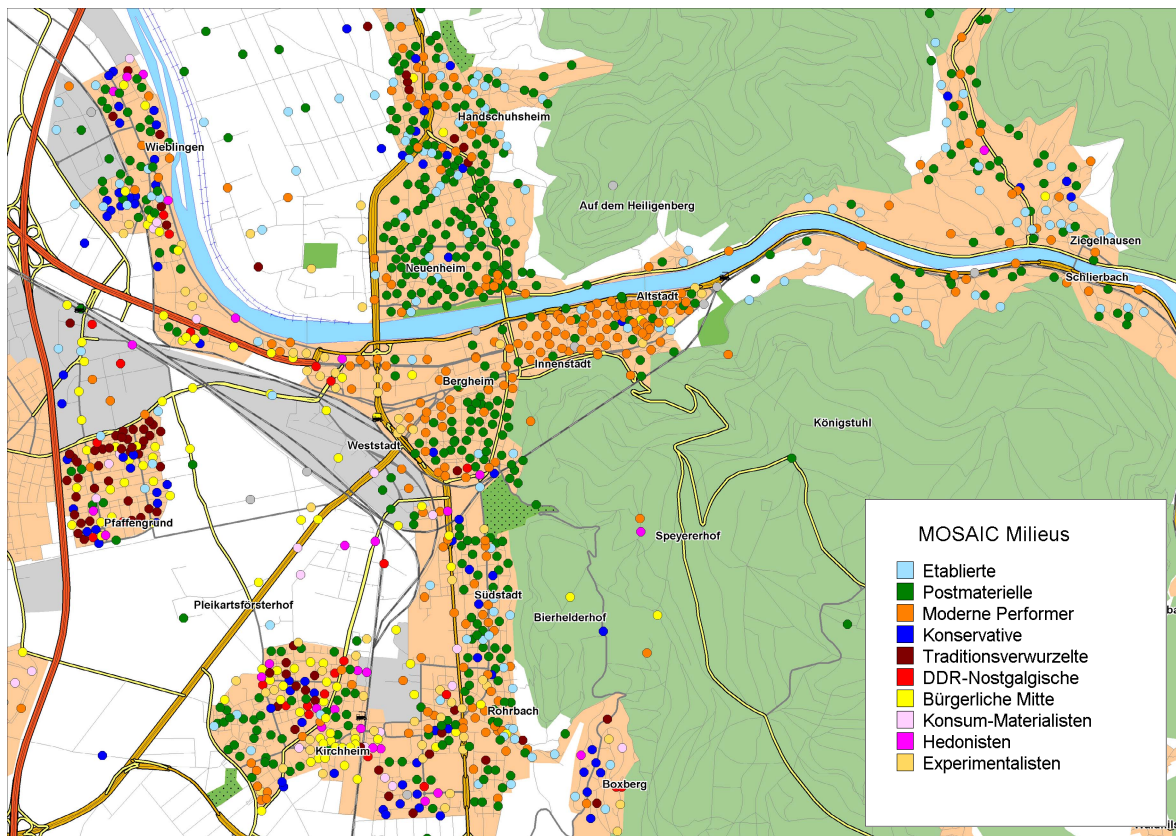


Abbildung 5: MOSAIC Milieus[®] auf Straßenabschnittsebene am Beispiel Heidelberg

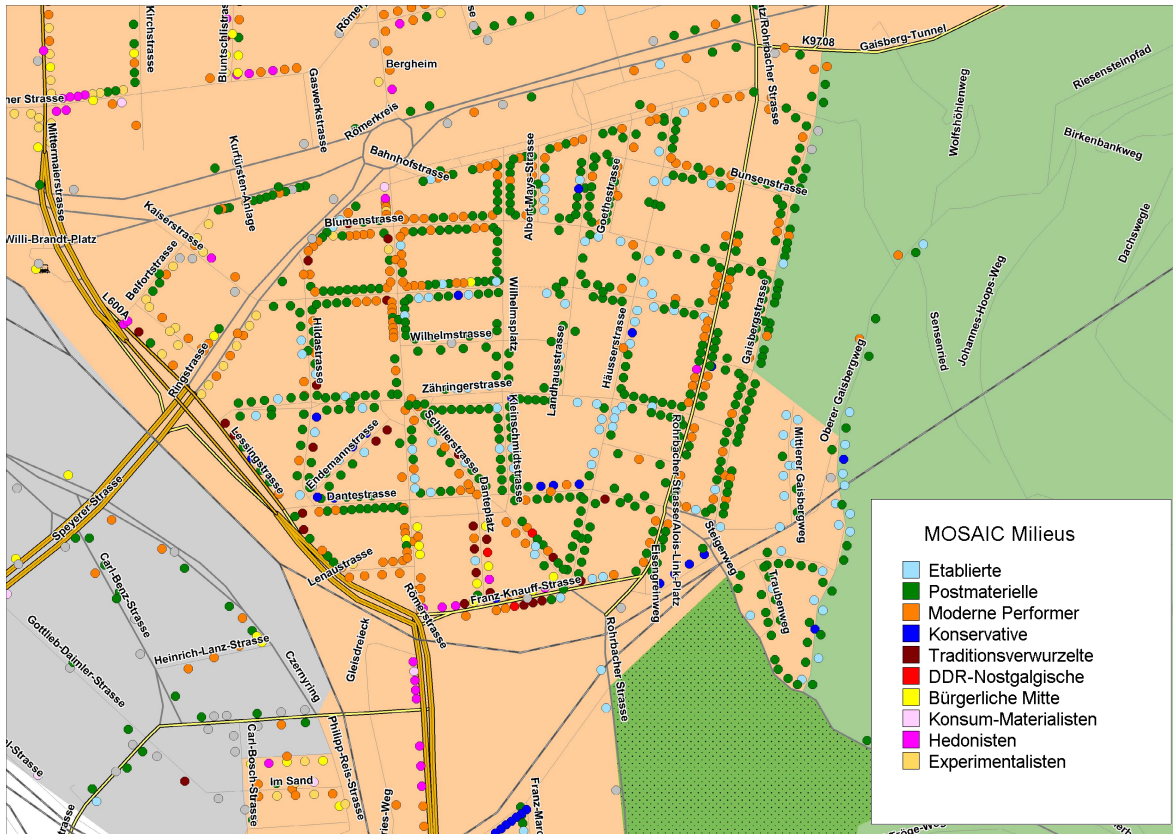


Abbildung 6: MOSAIC Milieus® auf Hausebene am Beispiel Heidelberg-Bergheim