Reduktion von Feldkosten durch automatisierte Clusterung von Adressen

Gabler, Siegfried

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

REDUKTION VON FELDKOSTEN DURCH AUTOMATISIERTE CLUSTERUNG VON ADRESSEN

SIEGFRIED GABLER

Liegen viele Adressen von zu befragenden Personen in einer Stadt vor, stellt sich für die Feldorganisation eines Umfrageinstituts die Aufgabe, die Adressen in Cluster zusammenzufassen, um Kosten zu sparen. Die Zahl der zu befragenden Personen in jedem Cluster sollte dabei möglichst gleich groß ist. Mit zunehmender Zahl der Adressen und Städte lohnt es sich, diese Clusterung zu automatisieren. Dies wird im folgenden am Beispiel des ALLBUS 1996 demonstriert.

If there are a lot of interview addresses in a city which are scattered over its area, it may be advantageous to cluster them to save field costs. In the ALLBUS 1994 this was done manually. The present paper describes a way to automatize this clustering process which was implemented in the ALLBUS 1996 survey.

1. Einleitung

Die Umstellung des Stichprobenplans von Random Route auf Einwohnermeldeamtsstichproben ab ALLBUS 1994 hatte zur Folge, daß zum einen weniger Orte als früher in die Auswahl kamen, innerhalb eines Ortes aber mehr Adressen zur Verfügung standen. Da innerhalb der Städte zumindest theoretisch alle Personen der ALLBUS-Gesamtheit dieselbe Wahrscheinlichkeit haben, ausgewählt zu werden, verteilen sich die Adressen über die ganze Stadt, während sie beim Random Route stark geklumpt sind. Dabei wurde in Großstädten das Vielfache an Adressen beschafft von dem, was eigentlich benötigt wurde. Der Grund dafür ist, daß bei einer festen Zahl von Adressen einer Stadt die Distanzen zwischen den Befragten - und damit die Kosten - mit der Größe der Stadt wachsen. Um die Kosten der Feldarbeit in Großstädten nicht zu groß werden zu lassen, eine dem Random Route gegenüber gewollte Streuung der Adressen über eine Stadt aber dennoch zu haben, wurden gestaffelt nach Einwohnerzahl einer Stadt zwei- oder dreimal soviele Adressen von den Einwohnermeldeämtern angefordert als dann später befragt werden sollten. Diese Adressen wurden beim ALLBUS 1994 zunächst auf Stadtplänen

mühsam von Hand gesucht, mit einem Etikett versehen und danach per Hand zu gleich großen Adressenklumpen mittels Schnüren zusammengefaßt. Diese Klumpen wurden schneckenförmig angeordnet und einige durch systematisches Zufallsverfahren ausgewählt und nach weiteren Veränderungen[1)] den Interviewern zur Verfügung gestellt. 1994 hat Infratest diese Arbeit selbst gemacht. Um die steigenden Kosten beim ALLBUS 1996 zu senken, übernahm ZUMA die Clusterung der Adressen selbst. Da beim ALLBUS 1996 Städte bereits ab 100.000 Einwohnern als Großstädte definiert wurden, betraf es 33 Großstädte im Westen und elf im Osten. Insgesamt mußten fast 9.000 Adressen behandelt werden. Bei ZUMA wurden daher Überlegungen zur Automatisierung der Clusterung angestellt. Das Ergebnis war, daß vom Empfang der Adressen einer Stadt bis zur Clusterung die Aufgabe in fünf Arbeitsschritten bewältigt wurde. Falls keine größeren Ungereimtheiten auftraten, war es nach einer Einarbeitung möglich, 120 oder 180 Adreß-Datensätze einer Stadt in etwa einer Stunde einlesen zu lassen, auf einem elektronischen Stadtplan zu markieren, die Adressen zu clustern und vier ausgewählte Cluster mit je 15 Adressen zu Infratest zurückzuschicken.

2. Beschreibung der fünf Arbeitsschritte

Um die Aufgabe der Clusterung von Adressen in einer Stadt zu automatisieren, verwendete ZUMA:

Editor und Datenbankprogramm,
CARDYTSM[2)] (Routenplaner),
GAUSS[3)] (Programmiersprache).

Die Adressen liegen normalerweise als Adressendateien im ASCII-Format vor mit folgendem Aufbau:

```
1-3 Pointnr (Gemeinde), 4-6 Idnummer pro Point, 7-26 Adreßzusatz,
27-66 Straße, 67-71 Hausnummer, 72-76 Postleitzahl, 77-116 Ort,
117-156 Ortsteil
```

Arbeitsschritt 1: Adressendatei aufbereiten

Die Datensätze einer Stadt werden mittels eines Editors oder Datenbankprogramms aus der Gesamtdatei herausgezogen und in einer Datei gespeichert. Falls notwendig, muß mittels einer Ersetzfunktion „S(s)traße“ in „S(s)tr.“ umgewandelt werden, sonst kennt später CARDYTSM die Straße nicht. Ebenso wird „P(p)latz“ zu „P(p)l.“. Falls in der Adressendatei keine Umlaute geschrieben wurden, sollte dies ebenfalls jetzt getan werden, falls dies keinen zu großen Aufwand bedeutet. Wenn dies nicht erfolgt, stoppt später CARDYTSM bei solchen Adressen, schlägt als neuen Namen aber meist gleich

den richtig Geschriebenen vor. Die für CARDYTSM benötigten Datensätze haben folgendes Aussehen:

- Die einzelnen Felder sind durch einen Delimiter (z.B. ;) getrennt sind. (Die Leerstellen können beliebig sein). Um die Daten für CARDYTSM lesbar zu machen, müssen die Datensätze dabei folgende Anordnung haben:

Beispiel:

❶;❷;❸;❹;❺;❻;❼;❽;❿ ...
`1;Maier;Josef;Dr.;BASF;16133;Musterstadt;Maximilianstr.;21`

Im einzelnen bedeutet

❶ Kundennummer (1 bis ...)
❷ Name (oder Identifikationsnummer)
❸ Vorname
❹ Titel
❺ Firma
❻ PLZ
❼ Ort
❽ Straße
❿ Hausnummer

Weitere Felder können noch angehängt werden, wie weiter unten gezeigt wird. Als Minimum müssen ❶,❷ und ❽ angegeben werden.

`1;Maier;;;;;;Maximilianstr.;`

Die Straße muß richtig geschrieben sein und darf die Hausnummer nicht enthalten. Im Editor werden in die ersten vier Zeilen noch Kommentare geschrieben, die von CARDYTSM überlesen werden.

Beispielhaft sieht die ASCII-Datei jetzt wie folgt aus:

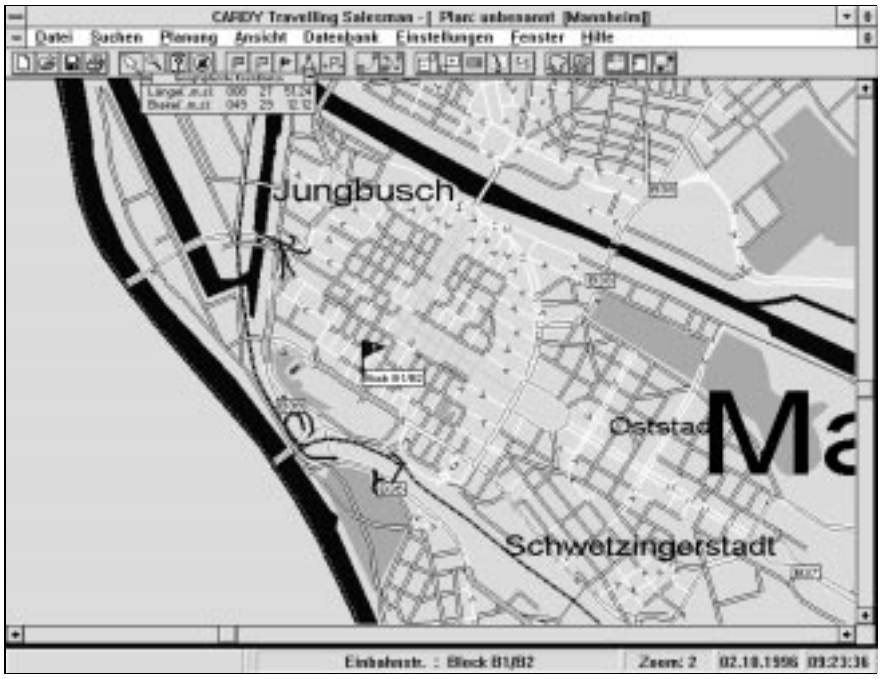
```
Umfrage XYZ
MUSTERSTADT 28.11.95
118 Adressen
**********************************************
1;Maier;;;;16133;Musterstadt;Maximilianstr.;21
```

↓ weitere 117 Datensätze

Arbeitsschritt 2: Verortung von Straßen

Um eine Clusterung der Adressen durchführen zu können, müssen Distanzen zwischen den Objekten (Adressen) bekannt sein. Dazu bieten sich die Entfernungen zwischen den Adressen an. Um diese berechnen zu können, bedarf es zunächst eines Programms, das Adressen aus einer Datei einlesen und ihnen ihre Weltkoordinaten zuordnen kann. Ein solches Programm haben wir mit CARDYTSM gefunden.

Abbildung 1: Kartenausschnitt aus Mannheim, in dem die Straße markiert ist, in der ZUMA sich befindet.



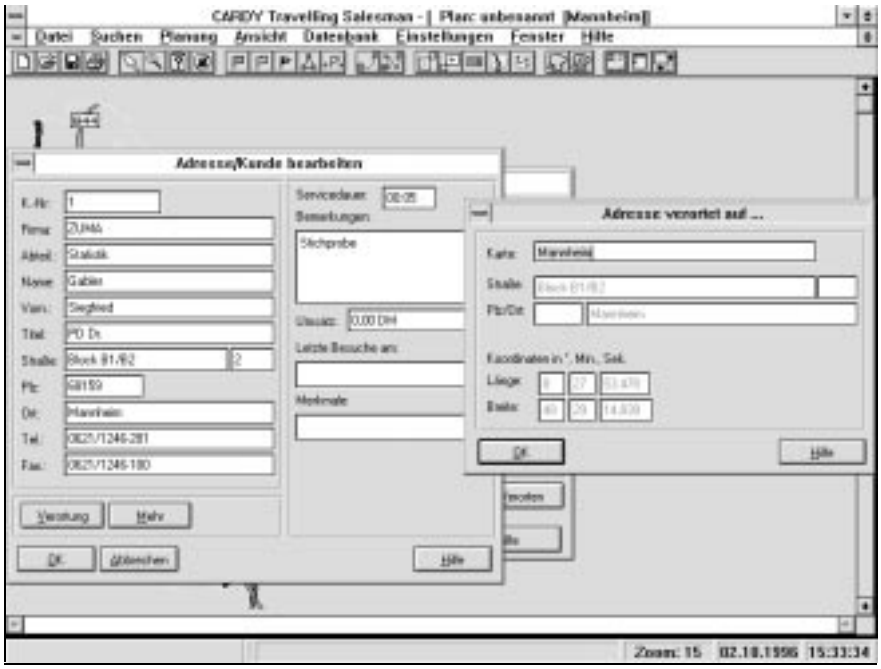
Man lädt aus den etwa 500 vorhandenen Stadtplänen die entsprechende Karte und holt über die IMPORT-Funktion die Daten in das Fenster. Jetzt kann man nochmals überprüfen, ob genausoviele Datensätze geladen wurden wie in der Datei stehen. Durch Anklicken des Knopfes VERORTEN verortet CARDYTSM alle Straßen.

Bei falscher Bezeichnung der Straßennamen sucht CARDYTSM den am ähnlichsten klingenden Namen und schlägt ihn vor. Änderungen sind jetzt möglich. Nach Verortung der Straßen können alle Adressen - etwa nach Postleitzahl sortiert - über die EXPORT-Funktion in eine ASCII-Datei gespeichert werden, deren ersten vier Zeilen in einem Editor für die Weiterverarbeitung zu löschen sind.

Arbeitsschritt 3: Vorbereitung für Clusterung und Kontrollschritt

Abbildung 2 zeigt im linken Teil noch einmal ein Beispiel dafür, welche möglichen Felder vom Anwender in der Input-Datenmaske bei CARDYTSM beschriftet werden können und im rechten Teil die Verortung der Adresse durch das Programm. Abbildung 1 liefert die Lage der Straße „Block B1/B2“ im Stadtplan von Mannheim, wobei zu beachten ist, daß der Name dieser Straße bei der Quadratestadt Mannheim nicht ungewöhnlich ist.

Abbildung 2: Beispiel für Input-Datenmaske bei CARDYTSM



Ein Export-Datensatz hat jetzt 35 Felder, unter denen sich auch die Breiten- und Längenangaben befinden. Beispiel für einen Datensatz einer Export-Datei:

1;Gabler;Siegfried;PD Dr.;ZUMA;68159;Mannheim;Block B1/B2;2;;
0621/1246-281;0621/1246-100;1413;4204;10:34;10:34;10:34; 10:34;
5;**Stichprobe**;0;0;0;0;1;16725119;49;29;14.8392;8;27;53.4784;0;-1;
Statistik

Die fettgedruckten Teile wurden als Input in die Daten-Maske von CARDYTSM eingegeben, die unterstrichenen Felder sind die berechneten Breiten- und Längengrade, die sonstigen Feldinhalte berechnet das Programm selbst.

Für die Clusterung werden aus einem exportierten Datensatz acht Felder herausgezogen, nämlich Name, geographische Breiten- und Längengrade sowie die Postleitzahl. Zuvor wurde noch ein Kontrollschritt eingefügt.

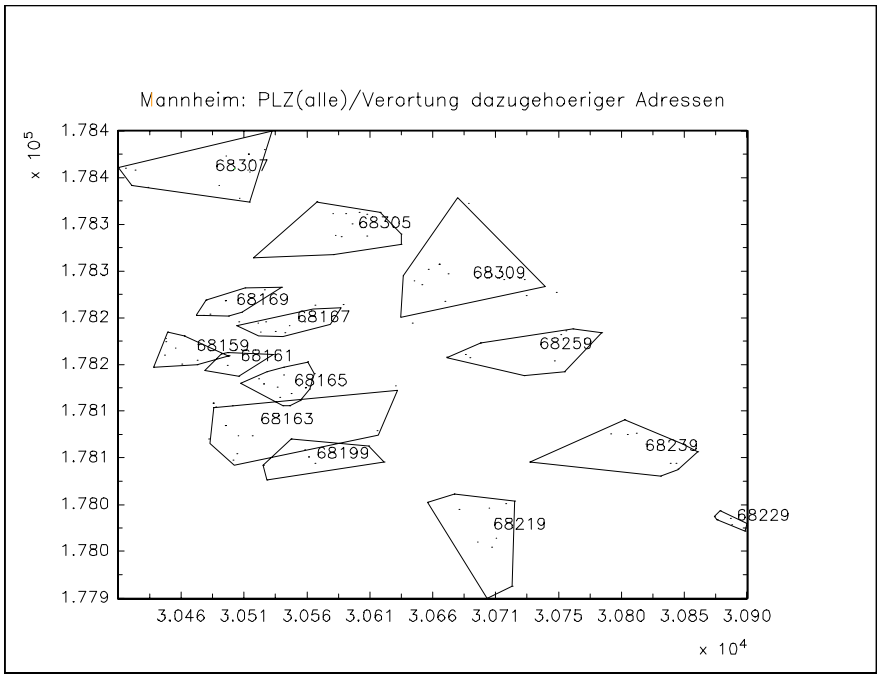
Die Verortung mittels CARDYTSM hat seine Tücken.

- Da eine Straße durch Angabe der Weltkoordinaten geocodiert ist, wird eigentlich nur ein Punkt der Straße festgehalten, der Mittelpunkt.
- Gibt es einen Straßennamen mehrmals in einer Stadt, was etwa in Berlin häufig vorkommt, sucht CARDYTSM eine dieser Straßen heraus. Häufig ist es jedoch nicht die auf die Adresse zutreffende. Eine einfache Möglichkeit, dies zu vermeiden, wäre gegeben, wenn CARDYTSM die Straßen über die Postleitzahlen identifizieren könnte, was aber leider nicht der Fall ist.
- Lange Straßen werden in Straßensegmente aufgeteilt, von denen CARDYTSM einen heraussucht.
- Manche Straßen gibt es in CARDYTSM nicht, insbesondere in den neuen Bundesländern sind die Straßennamen nicht immer auf dem neuesten Stand.
- Manchmal sind Straßen in CARDYTSM einfach falsch bezeichnet.

Um zu kontrollieren, ob Verortung und Postleitzahl zusammenpassen, wurde in GAUSS ein Programm geschrieben, das grafisch die verorteten Punkte mit zugehöriger Postleitzahl zeigt. Die zu einer Postleitzahl gehörenden Punkte werden mittels einer konvexen Hülle zusammengefaßt. Ein Vergleich mit dem Postleitzahlenbuch zeigt dann, ob die Verortung einer Straße im richtigen Bereich ist oder nicht. Wenn nicht, wird nachgeschaut, ob es denselben Straßennamen in der Stadt noch einmal gibt und neu in CARDYTSM verortet. Ähnliches gilt, wenn eine Straße mehrere Postleitzahlenbereiche quert und die automatische Verortung ein falsches Segment wählt. Wird eine Straße nicht gefunden oder ist deren Verortung offensichtlich falsch, wird einfach ein Punkt im Postleitzahlenbereich definiert. In kritischen Einzelfällen sollte zur Klärung auch ein

Stadtplan herangezogen werden. Im Durchschnitt mußten etwa fünf Prozent aller Adressen von Hand nachverortet werden.

Abbildung 3: Zuordnung von Postleitzahlen zu den verorteten Adressen am Beispiel Mannheim (x-Achse: Längengrade; y-Achse: Breitengrade)

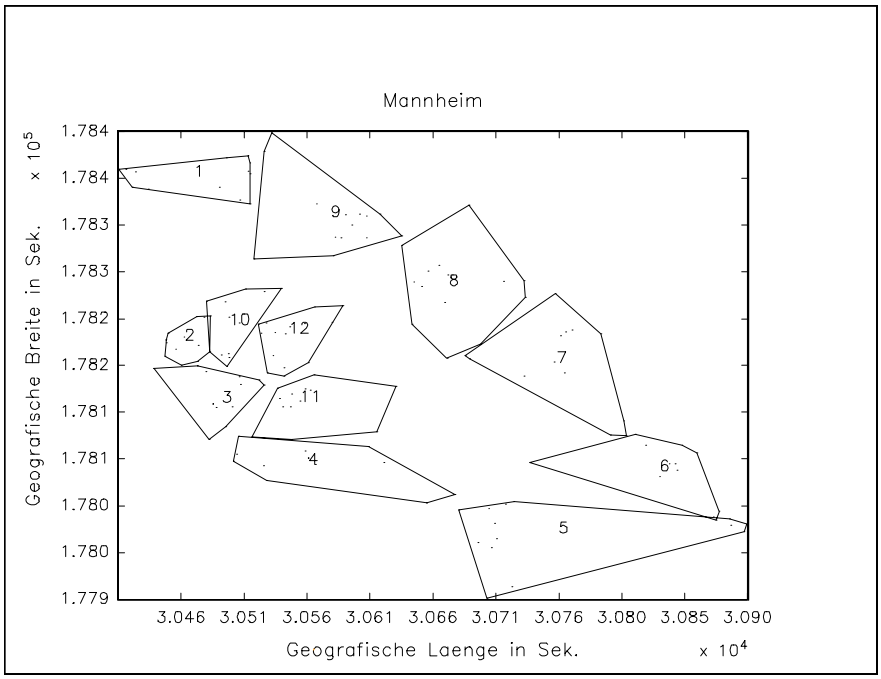


Arbeitsschritt 4: Clusterung der Adressen

Nach dem Kontrollschritt wird ein in GAUSS geschriebenes Programm gestartet, das die Anzahl der Adressen pro Cluster selbst berechnet, wobei sich die Clusterumfänge höchstens um 1 unterscheiden. Ausgehend von dem untersten Punkt in der linken Ecke faßt das Programm die diesem nächstliegenden Adressen zu einem Cluster zusammen und arbeitet sich so weiter, bis die Clusterzugehörigkeit aller Adressen gewährleistet ist. Die Cluster-Bilder können betrachtet werden. Jetzt ist es auch noch mittels einer Tauschprozedur möglich, Adressen von einem Cluster mit Adressen anderer Cluster zu vertauschen. Die Cluster werden danach im Programm schneckenförmig durchnumeriert und -

je nach Zahl der Adressen in einem Cluster - einige davon mittels systematischer Zufallsauswahl ausgewählt. Auf diese Weise wird verhindert, daß die ausgewählten Adressen zu dicht beieinander liegen und möglichst Randgebiete und Innenstadtregionen in der Stichprobe vertreten sind.

Abbildung 4: Alle zwölf schneckenförmig angeordneten Cluster in Mannheim



Alle Adressen werden in eine Datei geschrieben, mit Angabe der Clusternummer und der Kennzeichnung, ob die Adresse zu einem ausgewählten Cluster gehört oder nicht.

Beispiel für Musterstadt mit 118 Adressen.

```
Musterstadt                30. November 1995
Ausgewählte Cluster Nr   1   3   5   7  (Dritte Spalte = 1)
Lfd    CLu        Breitengrad     Längengrad
Nr     Nr          '  m  sek       '  m  sek
-----  ---  -    -------------    -------------
121001  5   1    51 22 37.8508    8 32 25.8199
↓ weitere 117 Datensätze
```

Außerdem wird eine Zusammenfassung der Clusterumfänge gegeben.

```
Die Clusterung ergab 8 Cluster
Cluster Nr.  Anzahl  ausgewaehlt
--------------------------------
   1           15       ja
   2           14       nein
   3           14       ja
   4           15       nein
   5           15       ja
   6           15       nein
   7           15       ja
   8           15       nein
              ----
              118
```

In einer Datei mit Bemerkungen sollten alle auftretenden Schwierigkeiten bei der Clusterung vermerkt werden.

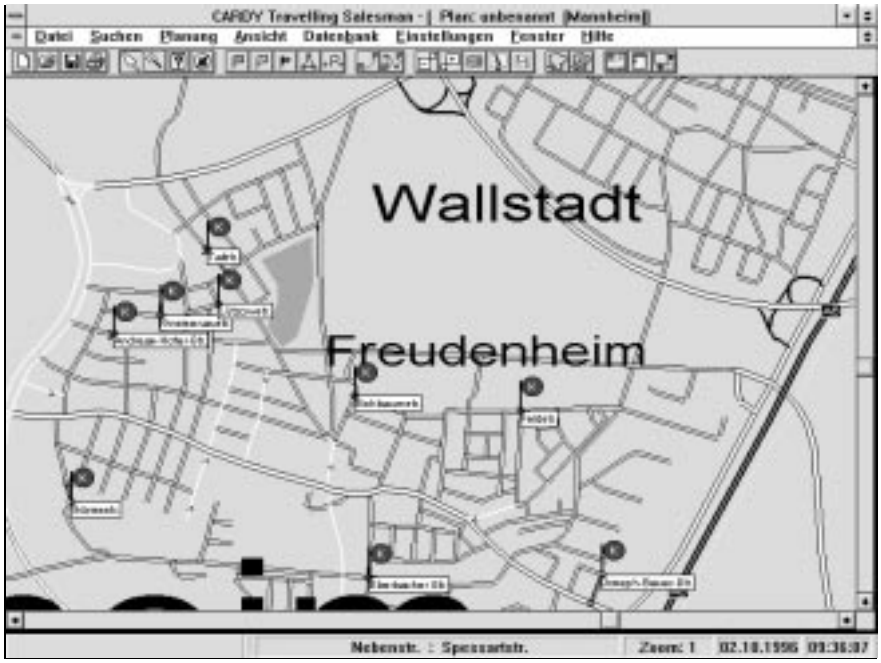
Arbeitsschritt 5: Grafik Output

Um für die einzelnen Cluster „Kundengrafiken“ aus CARDYTSM zu erhalten, müssen zunächst die Datensätze für die einzelnen Cluster in der Datenbank von CARDYTSM markiert oder für jedes ausgewählte Cluster einzelne Adressendateien erstellt werden. Danach können die Adressen, wie in Abbildung 5, angezeigt werden.

Zusammenfassung

Mit den oben aufgezeigten Arbeitsschritten hat ZUMA eine Möglichkeit, in relativ kurzer Zeit eine Fülle von Adressen in einer Stadt zu verorten und zu clustern. Schwierigkeiten liegen dabei nur in der Tatsache begründet, daß Straßennamen mehrfach in einer Stadt auftreten können und im nachhinein eine manuelle Nachbesserung notwendig werden kann. Interessenten für das in GAUSS geschriebene Clusterungsprogramm können sich an den Autor wenden.

Abbildung 5: Markierte Adressen der zu befragenden Personen



Bemerkung: Daß die Beschriftung nicht hundertprozentig richtig ist, zeigt sich in obiger Karte darin, daß CARDYTSM aus Feudenheim ein Freudenheim gemacht hat.

Anmerkungen

1) Es fand noch eine Anpassung an externe Größen wie Alter usw. statt. Näheres dazu ist im ALLBUS Methodenbericht zu lesen. Vgl. Koch, A./Gabler,S./Braun,M. 1994: Konzeption und Durchführung der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) 1994. ZUMA-Arbeitsbericht 94/11.

2) CARDY Travelling Salesman (1995). CARDY Karten Informationssysteme GmbH, Nobelstr. 3-5, 41189 Mönchengladbach.

3) GAUSS Version 3.2.13 (1995). Aptech Systems, Inc. 23804 South East Kent-Kangley Road, Mapple Valley, WA 98038 USA.