Entwicklung einer Familiendatei im Rahmen des Münchener Mikrodemographischen Analysesystems

In zwei Sonderheften des Jahrgangs 1976 dieser Schriftenreihe wurde die Einführung des Münchener Mikrodemographischen Analysesystems – MIDAS – vorgestellt. Die Arbeitsergebnisse, die mit der Einführung dieses Systems erzielt werden konnten, dienen in erster Linie dazu, dem steigenden Informationsbedarf der planenden Verwaltung an demographischen Daten in tiefer räumlicher Gliederung Rechnung tragen zu können. Aber auch zur Erforschung der demographischen Veränderungen, die sich vor allem im letzten Jahrzehnt in unseren Städten bemerkbar machten, war dieses System dienlich.

Im Rahmen der demographischen Arbeiten kommt in jüngster Zeit der Familienstatistik eine wachsende Bedeutung zu. 1970 boten für eine solche Statistik die Ergebnisse aus dem großen Zensus eine gute Grundlage. Da zu Beginn dieses Jahrzehnts der neue Zensustermin noch nicht festliegt und daher noch nicht abzusehen ist, wann Zählungsdaten für diesen demographischen Bereich zur Verfügung stehen werden, galt es, eine neue Basis zu finden. Es bot sich an, die Einwohnerdatei der Landeshauptstadt München entsprechend zu nutzen. Hierfür mußte ein neues Programm entwickelt werden. Einige Arbeitsergebnisse, die im Rahmen der neuen Familienstatistik erzielt werden konnten, sind bereits in Heft 11/1979 dieser Schriftenreihe mit dem Beitrag „Haushaltszahlen zwischen den Volkszählungen" und in Heft 12/1979 mit dem Beitrag „Mehrpersonenhaushalte in München und ihre Verteilung" veröffentlicht worden. Diese Veröffentlichungen haben über die Grenzen Münchens hinaus zu einem positiven Echo geführt und bei zahlreichen Fachkollegen den Wunsch hervorgerufen, mehr über die Erstellung der Familiendatei aus der Münchener Einwohnerdatei zu erfahren. Hierzu hat sich freundlicherweise der Leiter eines Organisations- und Programmierteams der Münchener EDV-Abteilung, Herr Dipl. Mathematiker Dr. Jung bereiterklärt, dessen Beitrag im folgenden veröffentlich wird. *Dr. Dh.*

Erstellung der Familiendatei aus der Einwohnerdatei

Vorbemerkung:

Die vorliegende Arbeit stellt die theoretische Grundlage für ein ADV-Programmsystem dar, womit aus der Einwohnerdatei der Landeshauptstadt München die für diverse planerische und demographische Untersuchungen benötigte Familiendatei gewonnen wird.

Die ersten Anregungen hierzu wurden im Zusammenhang mit stichprobenweisen Haushaltsbefragungen für Verkehrsplanungs- und Stadtteilsanierungszwecke bereits ab Anfang 1977 gegeben. Auf den dabei gewonnenen Erfahrungen aufbauend konnte vom Direktorium, Abt. Elektronische Datenverarbeitung, dem Wunsch des Statistischen Amtes entsprochen werden, die Familiendatei dem jeweiligen Bevölkerungsstand entsprechend vierteljährlich zu errichten.

Für die hierbei nötige Zusammenarbeit in den statistischen Anwendungsfragen danke ich Herrn E. Huss, der dieses Projekt von der Seite der Fachdienststelle her verantwortlich betreute. Für die Unterstützung im Zusammenhang mit der Einwohnerdatei danke ich Herrn W. Saller, dem Leiter des Organisations- und Programmierteams Einwohnerwesen. Meinem Mitarbeiter und Programmierteamleiter, Herrn F. Weber, spreche ich für Programmerstellungen und umfassende Laufzeitoptimierungen ebenfalls meinen Dank aus.

Problemstellung

In der Einwohnerdatei der Landeshauptstadt München ist im Rahmen eines Datenbanksystems für jede gemeldete Person jeweils ein Datensatz gespeichert, aus dem sowohl Angaben zu dieser selbst, als auch Hinweise zu deren Familienmitgliedern hervorgehen. Letzteres geschieht über sogenannte Ordnungsmerkmale, die in eindeutiger Weise jeder Person zugeteilt sind. In jedem Personendatensatz ist dieses Ordnungsmerkmal selbst und, sofern jeweils relevant, das des Ehegatten und das der Kinder enthalten. Eine logische Reihenfolge dieser Sätze im Sinne von Familienzusammengehörigkeiten besteht jedoch in der Einwohnerdatei nicht.

Um nun eine Familiendatei zu erhalten, liegt die Aufgabe darin, unter Zuhilfenahme der Ordnungsmerkmale Datensätze so zu konstruieren, daß in jedem bestimmte Einwohnerdaten aller zu jeweils einer Familie gehörigen Personen enthalten sind. Dies soll in möglichst geringer Verarbeitungszeit einer Großrechenanlage geschehen, so daß jährlich mehrmals jene Datei wirtschaftlich aus der Einwohnerdatei errichtet werden kann. Vor allem soll die Rechenzeit nur proportional zur Gesamtdatenmenge sein, und zwar auch dann, wenn Anhäufungen von Einwohnern unter gleichen Adressen, z.B. bei Anstalten oder Hochhäusern, vorkommen. Außerdem soll auf Zwischendateien verzichtet werden; die in Frage kommenden Datensätze sollen stets sequentiell gelesen bzw. geschrieben werden.

Die verkürzte Einwohnerdatei

Um das eigentliche Verfahren der Familienzusammenführung auf einen kleinstmöglichen Umfang an Daten zu beschränken, wird aus der Einwohnerdatei zunächst ein Auszug erstellt, der im folgenden als „verkürzte Einwohnerdatei" bezeichnet wird. Dessen Datensätze haben variable Länge und sind folgendermaßen aufgebaut*):

Bedeutung		Stellenzahl	binär (B) oder entpackt (E)
Satzlängenfeld		4	B
Straßen-Nummer	Adresse	5	E
Hausnummer	Adresse	4	E
Alpha-Zusatz	Adresse	2	E
Staatsangehörigkeit		3	E
Familienstand		2	E
Erwerbstätigkeit		1	E
Anzahl der folgenden Ordnungsmerkmale (OM)		2	E
OM der Person selbst		4	B
{OM des Ehegatten}		{4}	B
{OM des 1. Kindes}		{4}	B
{OM des 2. Kindes}		{4}	B
.		.	B
.		.	B
.		.	B
.		.	B
{OM des 20. Kindes}		{4}	B

*) Optimierung des Satzaufbaus i. Vergl. zu „Münchener Statistik" Jg. 1979, Heft 11, S. 258.

Bei dieser Darstellung sind, wie auch später, die eventuell enthaltenen Daten stets in geschweiften Klammern aufgeführt. Die Staatsangehörigkeit wird mit einer 3–5stelligen Schlüsselzahl angegeben, die bei Deutschen leer bleibt und bei Ausländern die erste ausländische Staatsangehörigkeit kennzeichnet. Auch der Familienstand und die Erwerbstätigkeit sind

zahlenmäßig verschlüsselt, wobei letztere durch den Lohnsteuerkartenbesitz simuliert werden muß, da weitere geeignete Angaben hierzu in der Einwohnerdatei fehlen. Bei verheirateten Personen sind die Ordnungsmerkmale der Kinder einschließlich Stiefkinder in den Datensätzen beider Elternteile angegeben, bei geschiedenen Personen jedoch nur die der jeweils leiblichen Kinder. Erwähnenswert ist ferner, daß obiger Datensatz auch dann vorkommt, wenn die Adresse der betreffenden Person nicht als Hauptwohnsitz gemeldet ist. In Ausnahmefällen erscheinen somit Personen sooft in der verkürzten Einwohnerdatei, wieviele Wohnsitze jene in München haben.

Bildung der Arbeitsspeicherbereiche

In Anwendung des Betriebssystem-Sortierprogrammes wird sodann die verkürzte Einwohnerdatei adressenmäßig in aufsteigende Reihenfolge gebracht. Die weitere Verarbeitung kann dadurch schrittweise erfolgen, und zwar so, daß in jedem im folgenden aufgezeigten Einzelschritt die Gesamtheit der Einwohnerdatensätze mit einer bestimmten Adresse zu den Familiendatensätzen zusammengeführt wird.

Hierbei werden zunächst die Sätze mit gleicher Adresse sequentiell in den Arbeitsspeicher eingelesen und in dieser Reihenfolge lückenlos auf die Felder NASTA, NAFAL, NAPKA und NAPKE verteilt. In NASTA sind die Staatsangehörigkeiten, in NAFAL zusammengefaßt die Familienstände und Erwerbstätigkeiten, sowie in NAPKA die eigenen Ordnungsmerkmale enthalten. Die eventuellen Ordnungsmerkmale von Ehegatten und Kindern sind unterschiedslos in NAPKE gespeichert, wobei ein begleitendes Adressierfeld NAEIN die zur Kennzeichnung der Zusammengehörigkeit erforderlichen Anfangs-Anordnungszahlen pro Einwohnersatz angibt. Für das Zusammenführungsverfahren werden außerdem die Indexfelder NAVER und NAVOR sowie ein logisches Feld LOEIN benötigt, die, wie auch NASTA, NAFAL und NAEIN, in Länge und inhaltsmäßiger Reihenfolge dem Feld NAPKA entsprechen. Die Felder NAVER und NAVOR erhalten zunächst die Anordnungszahlen der Einzelwerte in NAPKA, also die natürlichen Zahlen in lückenlos aufsteigender Reihenfolge. Das Feld LOEIN bekommt zu Anfang des Verfahrensschrittes die logischen Werte „Wahr" zugewiesen.

Anhand eines einfachen Beispiels aus der verkürzten Einwohnerdatei soll in Anhang 1 der Aufbau der obengenannten Felder veranschaulicht werden. Bei der Erläuterung des weiteren Verfahrensablaufes wird zum leichteren Verständnis auf dieses Beispiel noch wiederholt zurückgegriffen.

Erstellung der Indexketten

Das weitere Zusammenführungsverfahren vollzieht sich nun über die Felder NAPKA, NAPKE, NAEIN, NAVER, NAVOR und LOEIN. Dabei stellt sich zunächst allgemein die Frage, ob es zu einem in NAPKA enthaltenen Ordnungsmerkmal dort ein weiteres gibt, das ebenfalls in NAPKE in dem Bereich vorkommt, der durch den entsprechenden NAEIN-Wert und den darauffolgenden eingegrenzt wird. Hierzu wird ein weiteres Feld IVER geschaffen, das die Anordnungszahlen der Werte in NAPKA im aufsteigend geordneten Sinne enthält. Die genannte Frage läßt sich dann rechentechnisch jeweils durch Anwendung eines Bisektionsverfahrens auf IVER für jeden der in Frage kommenden NAPKE-Werte mit einem Vergleichs- und Rechenaufwand beantworten, der praktisch unabhängig von der Länge von NAPKA, also der Einwohnerzahl mit gleicher Adresse, ist. Der weitere Verfahrensablauf läßt sich nun durch die folgende Entscheidungstabelle beschreiben, die für jeden NAPKA-Wert einzeln anzuwenden ist.

		Bedeutung	R1	R2	R3	R4	R5	R6
Bedingungen	B 1	Ehegatte unter gleicher Adresse. .	J	J	J	N	N	N
	B 2	Kind unter gleicher Adresse	J	J	N	J	J	N
	B 3	Eheg. d. Kinder unter gl. Adresse .	J	N	–	J	N	–
Aktionen	A 1	Zuführung des Ehegatten	x	x	x			
	A 2	Zuführung des Kindes		x			x	
	A 3	Einm. Zuf. d. Kindes mit dessen Ehegatten				x		

Die in dieser Tabelle notwendige Unterscheidung zwischen Ehegatten und Kind ist dadurch möglich, daß im ersteren Fall das Ordnungsmerkmal der ursprünglich betrachteten Person wiederum in NAPKE an der zu deren Ehegatten gehörigen, in NAEIN angegebenen Stelle steht, im letzteren dagegen nicht. Jede Person kann nur einmal zugeführt werden.
Die Zuführung selbst geschieht mittels NAVER, NAVOR und LOEIN. Durch diese Felder werden Indexketten so definiert, daß zu jedem Ordnungsmerkmal in NAPKA die an gleicher Stelle in NAVER befindliche Anordnungszahl die Stelle des logisch nachfolgenden Ordnungsmerkmals in NAPKA bezeichnet und umgekehrt in NAVOR die des logisch vorausgehenden. Beim Durchlaufen einer Indexkette werden die Ordnungsmerkmale aller Personen, die zu einer ganz oder teilweise gebildeten Familie gehören, angezeigt. Jedes Ordnungsmerkmal ist einer und nur einer Indexkette zugehörig, wobei zu Beginn des Verfahrensschrittes gemäß obengenannter Anfangsdefinition von NAVER und NAVOR jede dieser Ketten sich jeweils über nur ein Ordnungsmerkmal erstreckt.
Bei jedem Zuführungsvorgang werden nun zwei Indexketten durch Anpassen der entsprechenden NAVER- und NAVOR-Werte vereinigt. Dabei erhält NAVER an der zur betrachteten Person gehörigen Stelle die Anordnungszahl der zuzuführenden Person in NAPKA zugewiesen und umgekehrt NAVOR. Die ursprünglichen Werte in NAVER und NAVOR werden in diesen Feldern überkreuzt an den Stellen zugewiesen, die durch sie selbst angegeben sind. LOEIN erhält an der Stelle der zuzuführenden Person den Wert „Falsch", so daß in jeder Indexkette stets einmal der dazugehörige LOEIN-Wert „Wahr" vorkommt. Eine solche Zuführung hat somit eine feste Zahl von Umspeicherungsvorgängen und ist in der Verarbeitungszeit unabhängig von der Einwohnerzahl pro Adresse.
Zur Veranschaulichung wird im Anhang 2 anhand des bereits eingeführten Beispiels der Indexkettenaufbau in den Feldern NAVER, NAVOR und LOEIN succesive aufgezeigt.

Aufbau der Familiendatei

Mit Hilfe der Indexketten läßt sich nun leicht aus den Arbeitsspeicherbereichen heraus sequentiell eine Familiendatei errichten, deren Sätze variable Länge und folgenden Aufbau haben*):

Bedeutung		Stellenzahl	binär (B) oder entpackt (E)
Satzlängenfeld		4	B
Straßen-Nummer	Adresse	5	E
Hausnummer	Adresse	4	E
Alpha-Zusatz	Adresse	2	E
Anzahl der Familienmitglieder(n)		2	E
1. Person			
Ordnungsmerkmal		4	B
Staatsangehörigkeit		3	E
Familienstand		2	E
Erwerbstätigkeit		1	E
2. Person			
dgl.		4	B
dgl.		3	E
dgl.		2	E
dgl.		1	E
.		.	
.		.	
.		.	
n. Person			
dgl.		4	B
dgl.		3	E
dgl.		2	E
dgl.		1	E

*) Optimierung des Satzaufbaus i. Vergl. zu „Münchener Statistik" Jg. 1979, Heft 11, S. 259.

Der Adressteil jedes Satzes ist stets bekannt, da bei jedem Schritt nur die zu der betreffenden Adresse gehörigen Einwohnersätze verarbeitet werden. Die weitere Satzerstellung geschieht durch paralleles Durchlaufen der Felder NAPKA, NASTA und NAFAL in der durch die Indexketten in NAVER festgelegten Reihenfolge. Jeder Indexkettendurchlauf beginnt an der Stelle, wo ein LOEIN-Wert „Wahr" ist und endet vor dessen Wiederkehr. Die Abarbeitungsreihenfolge der Indexketten richtet sich nach der Beschaffenheit von LOEIN, welches lückenlos durchlaufen wird und im Falle „Wahr" jeweils einen Einsprung in die betreffende Kette, im Falle „Falsch" den Übergang zum darauffolgenden LOEIN-Wert bewirkt. Dadurch ist ein nur einmaliges Abarbeiten jeder Indexkette gewährleistet, außerdem entfallen jegliche Suchvorgänge.

Für unser Anwendungsbeispiel ist der Inhalt der betreffenden Sätze der Familiendatei in Anhang 3 wiedergegeben. Wie man erkennt, handelt es sich hier um eine Dreigenerationen-Familie und eine Einzelperson.

Zu erwähnen ist noch, daß Personen, die in München mehrere Wohnsitze haben, auf Grund ihres mehrfachen Erscheinens in der verkürzten Einwohnerdatei ebensooft in den Sätzen der Familiendatei vorkommen. Dies bedeutet, daß die Familiendatei in gewissem Maße haushaltsbezogen zu verstehen und die Gesamtzahl der darin aufgeführten Personen geringfügig höher als die Einwohnerzahl Münchens ist. Dagegen können Wohngemeinschaften nicht als Haushalte dieser Art ausgewiesen werden, da hier auf Grund fehlender Hinweisdaten in der Einwohnerdatei keine Zusammenführung möglich ist.

Die zusammengefaßte und die zugeordnete Familiendatei

Da im allgemeinen eine Weitergabe der Familiendatei aus Datenschutzgründen nicht möglich ist, werden aus ihr zwei weiter anonymisierte Varianten abgeleitet.

Die erstere beinhaltet zu jeder Adresse die jeweilige Anzahl der Ein- bis Fünf- oder Mehrpersonen-Haushalte. Diese zusammengefaßte Familiendatei wird für Planungszwecke verwendet und läßt sich leicht aus der Familiendatei, die nach Adressen geordnet ist, sequentiell gewinnen.

Bei der letzteren Variante werden in Hinblick auf planerische und demographische Informationsbeschaffungen die etwas abgekürzten Adressen in den Familiensätzen um die dazugehörigen kleinräumigen Gliederungsmerkmale, wie Stadtbezirk, -teil, Viertel, Block und die jeweiligen Gauß-Krüger-Koordinaten ergänzt. Diese Zuordnung läßt sich jedoch auf Grund des ungleichen Aktualitätsstandes der dazu benötigten Datei nicht immer eindeutig vollziehen, weswegen solche Fälle durch einen gesonderten Eintrag gekennzeichnet werden. Außerdem erhalten die Sätze dieser zugeordneten Familiendatei anstelle des jeweiligen Ordnungsmerkmales das daraus zu ermittelnde Geschlecht und Geburtsjahr, jeweils in entpackter Schreibweise. Der Satzaufbau stellt sich somit folgendermaßen dar:

Bedeutung		Stellenzahl	binär (B) oder entpackt (E)
Satzlängenfeld		4	B
Blocknummer		6	
Straßen-Nummer	Adresse	5	
Hausnummer	Adresse	3	
Alpha-Zusatz	Adresse	1	
Koord.-Rechtswert		7	
Koord.-Hochwert		7	
Kennung ('*') für nicht eindeutige Zuordnung		1	
Anzahl der Familienmitglieder(n)		2	
1. Person			
Geschlecht		1	
Geburtsjahr		4	
Staatsangehörigkeit		3	
Familienstand		2	
Erwerbstätigkeit		1	E
2. Person	dgl.	1 4 3 2 1	
. . .		. . .	
n. Person	dgl.	1 4 3 2 1	

Angaben zu den Rechenprogrammen

Die Herstellung der verkürzten Einwohnerdatei erfolgt mit dem maschinenorientierten Programm STAFAD sequentiell aus den Sicherungs-Bandabzügen der Einwohnerdatei, die selbst in Anwendung des Datenbankverwaltungssystems ADABAS on-line betrieben wird. Hierbei wird im Falle der Stadt München mit ca. 1,3 Mio. Einwohnern eine Maschinenlaufzeit von ca. 2,5 Stunden an einer Siemens-7760-Anlage benötigt. Das resultierende Datenvolumen läßt sich auf nur einem Magnetband bei einer Beschreibungsdichte von 1600 BpI unterbringen, was auch jeweils für die folgenden Dateien gilt. Diese verkürzte Einwohnerdatei wird dann mit dem Systemprogramm SORT nach Adressen sortiert, wofür ca. 0,5 Stunden Rechenzeit erforderlich sind.

Die eigentliche Erstellung der Familiendatei wird mit dem Programm STAFAO vorgenommen, das in der Programmiersprache FORTRAN IV geschrieben ist. Trotz bewußter Vermeidung maschinen- oder systemorientierter Programmiermethoden erfordert dieses Programm lediglich eine CPU-Zeit von ca. 1,5 Stunden an einer Siemens-7748-Rechenanlage, wobei festgestellt wurde, daß über 80% dieser Zeit allein auf die Lese- und Schreibvorgänge entfallen. Der Arbeitsspeicherbedarf dieses ca. 220 Anweisungen umfassenden Programmes beträgt ca. 200 K-Bytes, wodurch eine maximale Einwohnerzahl von 2000 für eine Adresse verarbeitet werden kann.

Die zusammengefaßte und die zugeordnete Familiendatei wird mit den FORTRAN-IV-Programmen STAFAI und STAFAU gewonnen, wofür CPU-Zeiten von ca. 0,5 bzw. 1,5 Stunden benötigt werden. Auch hier wurde auf eine rein problemorientierte Programmierung geachtet.

Anhang 1

Beispielsweise entfallen in der verkürzten Einwohnerdatei 5 Sätze auf die Adresse ADRK, nämlich

Stelle	Bedeutung	Satz 1	Satz 2	Satz 3	Satz 4	Satz 5
1– 4	Satzlänge	35	27	27	35	35
5–15	Adresse	ADRK	ADRK	ADRK	ADRK	ADRK
16–18	Staatsangehörigkeit	STA 1	STA 2	STA 3	STA 4	STA 5
19–21	Familienstand, Erwerbstätigkeit	FAL 1	FAL 2	FAL 3	FAL 4	FAL 5
22–23	Anz. d. Ordnungsmerkm. (OM)	3	1	1	3	3
24–27	OM der Person	OM 1	OM 2	OM 3	OM 4	OM 5
28–31	OM des Ehegatten	OM 4			OM 1	OM 6
32–35	OM des 1. Kindes	OM 2			OM 2	OM 4

Dann haben die Felder	NASTA	NAFAL	NAPKA	NAPKE	NAEIN
	STA 1	FAL 1	OM 1	OM 4	1
	STA 2	FAL 2	OM 2	OM 2	3
folgenden Inhalt:	STA 3	FAL 3	OM 3	OM 1	3
	STA 4	FAL 4	OM 4	OM 2	3
	STA 5	FAL 5	OM 5	OM 6	5
				OM 4	

Anhang 2

Die Indexketten werden successive wie folgt erstellt:

	NAVER	NAVOR	LOEIN	Indexketten-verlauf
Anfangszustand	1 2 3 4 5	1 2 3 4 5	W W W W W	1 2 3 4 5
Zustand nach Abarbeitung von	4 2 3 1 5	4 2 3 1 5	W W W F W	1 2 3 4 5
Satz 1	2 4 3 1 5	4 1 3 2 5	W F W F W	1 2 3 4 5
Satz 5	2 4 3 5 1	5 1 3 2 4	F F W F W	1 2 3 4 5

Anhang 3

Die betreffenden Sätze der Familiendatei haben dann folgenden Inhalt:

Stelle	Bedeutung	Satz 1	Satz 2
1– 4	Satzlänge	27	57
5–15	Adresse	ADRK	ADRK
16–17	Anzahl der Familienmitglieder	1	4
18–27	1. Person:		
	Ordnungsmerkmal	OM 3	OM 5
	Staatsangehörigkeit	STA 3	STA 5
	Familienstand, Erwerbstätigkeit	FAL 3	FAL 5
28–37	2. Person:		
	dgl.		OM 1 STA 1 FAL 1
38–47	3. Person:		
	dgl.		OM 2 STA 2 FAL 2
48–57	4. Person:		
	dgl.		OM 4 STA 4 FAL 4