Omgaan met ontbrekende gegevens in statistische databases: Multiple imputatie in HERMES

Stef van Buuren[1]
Erik van Mulligen[2,3]
Jaap Brand[1,2]

Samenvatting

Dit artikel gaat in op enkele problemen en oplossingen ten aanzien van ontbrekende gegevens in statistische databases. Een veelbelovende oplossing, multiple imputatie, kent een tweetal hindernissen die een routinematige toepassing in de praktijk bemoeilijken. Een eerste is dat dezelfde gecompleteerde data idealiter geschikt moeten zijn voor meerdere analyses. Imputaties moeten daarom zoveel mogelijk gekozen worden onafhankelijk van modellen die in een later stadium op de data zullen worden toegepast. Dit kan worden bereikt door imputaties te vinden die de struktuur in de data, alsmede de onzekerheid rond deze struktuur, extrapoleren. Een tweede probleem is het efficiënte gebruik van complete-data methoden. Het HERMES werkstation biedt een architectuur waarin bestaande statistische programmatuur wordt ingebed onder een client-server model en vormt daardoor een geschikte omgeving voor de implementatie van multiple imputatie.

[1] Vakg. Statistiek, TNO Preventie en Gezondheid, Postbus 2215, 2301 CE, Leiden, buuren@nipg.tno.nl
[2] Vakg. Medische Informatica, Erasmus Universiteit, Postbus 1738, 3000 DR Rotterdam, brand@mi.fgg.eur.nl
[3] Academisch Ziekenhuis Dijkzigt, Postbus 55, Rotterdam, vanmulligen@mi.fgg.eur.nl

1 Introduktie

Het ontbreken van informatie is vaak een lastig praktijkprobleem bij het verzamelen en analyseren van onderzoeksgegevens. Onvolledige gegevens kennen veel oorzaken: een respondent vult niet alle items in van een vragenlijst, meetapparatuur kan defect raken, proefdieren kunnen onbedoeld overlijden, etc. Enkele problemen die direkt met ontbrekende data samenhangen zijn: 1) specifieke subgroepen kunnen uitvallen waardoor de steekproef niet meer representatief is, 2) er is minder informatie beschikbaar dan was voorzien, resulterend in geringere power bij het toetsen van hypothesen, en 3) standaard methoden voor complete data zijn vaak niet meer direct bruikbaar voor incomplete data. Ondanks grote inspanningen ten tijde van de dataverzameling zijn missing data onlosmakelijk met empirisch onderzoek verbonden. Geconfronteerd met incomplete data kiezen vele onderzoekers voor eenvoudige oplossingen zoals het weglaten van alle onvolledige cases of het imputeren (invullen) van het gemiddelde van het geobserveerde deel van de data. Veel statistische programmatuur doet dit soort zaken automatisch. Deze methoden zijn in veel gevallen voor kritiek vatbaar en kunnen de conclusies die uit de data worden getrokken ernstig ondermijnen (Little en Rubin, 1987, blz. 4, 44). Er bestaat duidelijk behoefte aan een gemakkelijke, algemeen toepasbare en statistisch correcte methode voor de behandeling van ontbrekende gegevens.

Een voorbeeld waar zo'n methode zijn nut kan bewijzen is het HERMES medische werkstation (van Mulligen, 1993). HERMES kent twee hoofdfuncties. Ten eerste kan het systeem automatisch een database samenstellen door cases te selecteren uit fysiek verschillende databestanden binnen hetzelfde ziekenhuis. Delen worden gecombineerd tot één bestand door cases op een gezamenlijke sleutel te matchen. Ten tweede kan het systeem gebruikt worden voor de statistische analyse van de data, zoals bijvoorbeeld beschrijvende statistiek of survival analyse. HERMES kan hierbij gebruik maken van bestaande software zoals BMDP, SPSS of SAS. Het zal duidelijk zijn dat het combineren van gegevens uit verschillende bronnen gemakkelijk kan leiden tot zeer onvolledige bestanden. Het HERMES medisch werkstation is ontworpen voor gebruik door artsen, die doorgaans geen uitgebreide opleiding hebben genoten in de database theorie en de statistiek. Gebruikersgemak en statistische correctheid zijn daarom van het allergrootste belang. Gezien deze uitgangspunten moeten problemen met onvolledige data dan ook zoveel mogelijk achter de schermen worden opgelost.

Dit artikel presenteert een algemene aanpak voor het behandelen van missing data problemen in statistische databases. Hoewel de principes breder toepasbaar zijn zullen we ons beperken tot item nonresponse, d.w.z. tot situaties waarin de respondent één of meer items niet beantwoordt. De methodologie is gebaseerd op multiple imputatie (Rubin, 1987). Gebruikmakend van een client-server architectuur is het mogelijk multiple imputatie op een transparante wijze te implementeren. Uiteraard moet de gebruiker zich ervan bewust zijn dat intern multiple imputatie

wordt toegepast, en zal hij of zij toegang moeten hebben tot elementaire imputatiematen. Daarnaast zal een expert mogelijk specifieke imputatieparameters willen instellen. Anders dan nu zal de gebruiker zich echter nauwelijks bezig hoeven te houden met de (soms gecompliceerde) details van de techniek. Het systeem dat we voor ogen hebben stelt de gebruiker in staat multiple imputatie toe te passen op een routinematig manier zonder veel extra werk. Het is het eerste voorbeeld van een dergelijk systeem.

Hieronder bespreken we kort enkele veelgebruikte methoden voor het analyseren van incomplete gegevens. We geven aan hoe multiple imputatie tegemoet komt aan de belangrijkste bezwaren van deze methoden. De tekst vat kort het idee achter multiple imputatie samen en geeft de belangrijke keuzes aan. Vervolgens laten we zien hoe multiple imputatie geïmplementeerd kan worden binnen het HERMES werkstation. Tenslotte gaan we in op de huidige status van het projekt.

2 Methoden voor incomplete gegevens

De eenvoudigste incomplete data methode is verwijdering van alle records met één of meer ontbrekende waarden. Dit wordt wel 'complete-case analysis' of 'listwise deletion' genoemd. Belangrijkste voordeel van deze methode is haar eenvoud. Groot nadeel is het potentiële verlies aan kostbaar verzamelde data. Ernstiger nog is dat schattingen onzuiver worden indien de incomplete observaties systematisch afwijken van de complete. Een iets subtielere methode is alle observaties in de analyse te betrekken waarin de doelvariabele(n) geobserveerd is. Deze aanpak staat bekend als 'available-case analysis' of 'pairwise deletion'. Alhoewel de aanwezige informatie efficiënter wordt gebruikt leidt ook deze methode in een aantal gevallen tot onzuivere schatters. Tevens zijn covariantie matrices niet noodzakelijkerwijs semi-positief definiet, hetgeen een noodzakelijke voorwaarde schendt voor het gebruik van veel multivariate technieken, waaronder regressie analyse. Over het algemeen zijn verwijderings-methoden dan ook onbevredigend.

Een tweede methode is het imputeren (invullen) van onbekende cellen door 'redelijke' waarden, waardoor een complete dataset ontstaat. Groot voordeel is dat bestaande complete-data analysetechnieken kunnen worden gebruikt. Een veel gebruikte (maar dikwijls slechte) imputatiemethode is het invullen van het gemiddelde van de geobserveerde data. Er zijn echter ook tal van verfijndere imputatietechnieken ontwikkeld. Zie hiervoor hoofdstuk 4 van Little en Rubin (1987). Geen van deze methoden is echter in staat de onzekerheid van de imputatie weer te geven. De complete-data analyse behandelt de geïmputeerde data als bekend. Rubin (1987) laat zien dat hierdoor de precisie van de schatters wordt overschat, betrouwbaarheidsintervallen te kort worden en correlaties kunnen worden vertekend. Enkelvoudige imputatiemethoden kunnen zodoende leiden tot onjuiste conclusies.

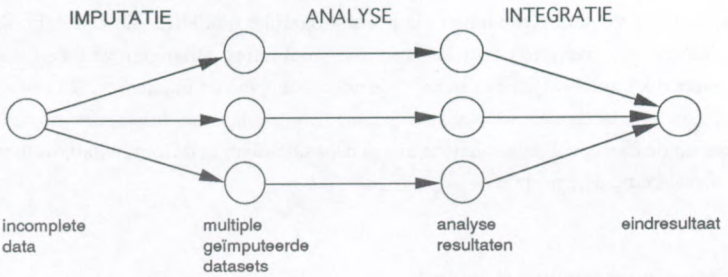
Een derde strategie is aanpassing van de statistische analyse. Een gebruikelijke startassumptie is de veronderstelling dat missing data Missing at Random (MAR) zijn. Dit houdt in dat eventuele verschillen in responskansen afhangen van geobserveerde data. Het volgende voorbeeld verklaart het begrip nader. Stel dat in een ongewogen random huishoudsteekproef een deel van de respondenten niet bereid is het inkomen op te geven. Indien de kans op nonresponse voor iedereen gelijk is dan spreken we van een Missing at Random (MAR) nonresponse mechanisme. Het is echter bekend dat inkomensvragen slechter worden beantwoord naarmate het inkomen stijgt. In dit geval hangt de kans op nonresponse mede af van het onbekende inkomen en zijn de data daardoor Missing Not at Random (MNAR). Stel dat naast inkomen ook is gevraagd naar bijvoorbeeld het aantal auto's binnen het huishouden en dat deze vraag door iedereen is beantwoord. Als nu binnen klassen van huishoudens met hetzelfde aantal auto's gelijke responskansen gelden dan is het mechanisme weer MAR. Verschillen in nonresponse kansen kunnen dus worden verklaard door bekende data, hier het aantal auto's per huishouden. Indien dat niet zo is, dan zijn de data MNAR. Little en Rubin (1987) hebben een theoretisch elegant framework ontworpen waarin modellen voor de nonresponse en voor de steekproef worden gecombineerd tot één likelihoodfunktie. Indien de missing data MAR zijn dan verdwijnt het missing data mechanisme uit de likelihood en heeft men aan de geobserveerde data voldoende. Onder een juist nonresponse model levert deze aanpak zuivere maximum likelihood schatters op, maar de uitvoering ervan vereist dikwijls veel statistische expertise. In een aantal gevallen is het onvermijdelijk speciale software te ontwikkelen omdat de wiskunde onwerkbaar is.

Het is gewenst een methode te vinden die a) de mogelijkheid biedt om bestaande software voor complete data te gebruiken, en b) het onder juiste aannamen mogelijk maakt om te komen tot zuivere parameterschattingen en een correcte weergave van de onzekerheid omtrent de missing data. Methoden 1 en 2 voldoen aan criterium a), maar niet aan b). Methode 3 voldoet wel aan b), maar niet aan a). Rubin heeft een alternatief ontwikkeld, multiple imputatie, dat aan beide criteria voldoet. Hieronder gaan we dieper in op de techniek.

3 Multiple imputatie van ontbrekende data

3.1 Modelgestuurde multiple imputatie

Het idee achter multiple imputatie is dat voor elke ontbrekende waarde niet één maar meerdere (m) imputaties worden gezocht. De spreiding tussen de m imputaties drukt de mate van onzekerheid van de imputatie uit. Onder een MAR mechanisme hangt deze spreiding af van de mate waarin de onbekende waarde uit de overige data voorspeld kan worden. Indien dit slecht gaat dan wordt de variatie tussen de m imputaties groot. Multiple imputatie creëert m



Figuur 1: Schematische weerwage van multiple imputatie met m = 3

gecompleteerde datasets die elk met behulp van dezelfde standaard complete-data methode wordt geanalyseerd. De m resulterende analyses worden vervolgens gecombineerd in een eindresultaat. Hiervoor bestaat een simpele procedure. Figuur 1 bevat een stroomschema van de voornaamste acties in multiple imputatie.

Rubin (1987) toont aan dat wanneer het complete data model bij absentie van nonresponse juiste schatters oplevert en wanneer de imputatieprocedure 'proper' is dan levert multiple imputatie zuivere punt- en intervalschattingen als m oneindig nadert. In 'proper' procedures wordt de variantie van de imputaties mede bepaald door de variantie van de parameters van het imputatiemodel. In de praktijk blijkt dat $m = 5$ vaak al voldoende is.

Het automatisch en routinematig toepassen van de techniek is niet zonder hindernissen. Ten eerste is de architectuur van conventionele statistische programmatuur niet echt geschikt voor implementatie. Met enig kunst- en vliegwerk is soms wel multiple imputatie uitvoerbaar, maar dat is verre van automatisch. Paragraaf 4 beschrijft een geïntegreerde omgeving die zich beter leent om bestaande statistische software in te passen. Ten tweede is het genereren van geschikte imputaties geen triviale taak. Het grootste obstakel hierbij vormt het afleiden van de verdeling van de missing data gegeven de bekende data en gegeven het nonresponse mechanisme. Deze verdeling wordt wel aangeduid als de predictieve verdeling. Imputaties worden uit de predictieve verdeling getrokken. Hieronder gaan we nader in op factoren die van invloed zijn op de bepaling van de predictieve verdeling.

Multiple imputatie vereist een specificatie van het steekproefmodel (het wetenschappelijk model) en een specificatie van het missing data mechanisme waarvan verondersteld wordt dat deze ten grondslag ligt aan het optreden van nonresponse (Little en Rubin, 1987). In de praktijk is het ware steekproefmodel zelden of nooit bekend, hetgeen het genereren van geschikte imputaties bemoeilijkt. In de oorspronkelijke formulering van multiple imputatie zijn imputaties zuiver modelgestuurd. Dit houdt in dat we iedere keer dat we iets aan het steekproefmodel veranderen de data opnieuw moeten imputeren onder het nieuwe model. Buiten het feit dat dit buitengewoon onpraktisch is, is het niet realistisch te veronderstellen dat het mogelijk is een

juiste predictieve verdeling te vinden voor alle mogelijke modellen die een gebruiker zou kunnen hanteren. Bovendien kan het feit dat imputaties afhangen van een specifiek wetenschappelijk model twijfel zaaien t.a.v. de neutraliteit van de imputaties. Dit laatste is een cruciale factor in de acceptatie van welke imputatiemethode dan ook. Imputaties moeten niet de conclusies uit de daaropvolgende analyse in een bepaalde richting duwen (tenzij, wellicht, deze richting met zekerheid de juiste is).

3.2 Datagestuurde multiple imputatie

Bovenstaande overwegingen hebben geleid tot een iets andere benadering t.a.v. de definitie van predictieve verdelingen. Het concept wordt 'mindless imputation' genoemd, een term die door Rubin is geopperd (persoonlijke medeling, 1993). We vertalen het hier met 'datagestuurde imputatie'. In datagestuurde aanpak hoeft het model dat wordt gebruikt om imputaties te creeren niet hetzelfde te zijn hoeft als het steekproefmodel. De term 'mindless' refereert aan de mogelijkheid dat het imputatiemodel vanuit wetenschappelijk oogpunt bezien onzinnig, irrelevant en overgeparametriseerd kan lijken. Dit is echter precies wat de aanpak inhoudt: steekproef-modellen (dat zijn modellen waarin de onderzoeker alle kennis over de data samenvat) die in een later stadium getest moeten worden (omdat het niet zeker is dat ze correct zijn) spelen geen essentiële rol tijdens imputatie.

In plaats hiervan stelt de datagestuurde benadering dat imputaties gezocht moeten worden die de struktuur van de data, alsmede de onzekerheid over deze struktuur, bewaren (zie Van Buuren, van Rijckevorsel en Rubin, 1993). Ter illustratie, stel dat de relatie tussen twee empirische variabelen ongeveer kwadratisch verloopt en dat het nonresponse mechanisme MAR is. Datagestuurde imputaties voor de y-variabele gegeven een x-waarde worden dan gegenereerd zodanig dat dezelfde kwadratische relatie ook behouden blijft binnen de gecompleteerde data, zonder dat de imputatiemethode gebruik maakt van de aanname dat deze variabelen een kwadratische relatie hebben. Een voorbeeld van een mindless methode is het fitten van een rijk en flexibel niet-lineair regressiemodel te fitten, op x te conditioneren, en aan de voorspelde y-waarde een hoeveelheid ruis toe te voegen die overeenkomt met de onzekerheid van de werkelijke y gegeven x. Datagestuurde imputaties zijn zoveel mogelijk ongevoelig en neutraal ten aanzien van de vervolganalyse. We gaan ervan uit dat dit bereikt kan worden door imputaties te vinden die in zeker opzicht dicht bij de data (en hun onzekerheid) staan.

Rubin en Schafer (1990) hebben een imputatiemethode voorgesteld die een geschikt vertrekpunt is voor de ontwikkeling van datagestuurde imputatiemethoden. In de Rubin-Schafer methode wordt veronderstelt dat observaties uit een multivariate normaalverdeling zijn getrokken en dat het nonresponse mechanisme MAR is. Imputaties worden gevonden door iedere variabele lineair uit alle andere variabelen te voorspellen. Deze methode van wisselende regressies is een

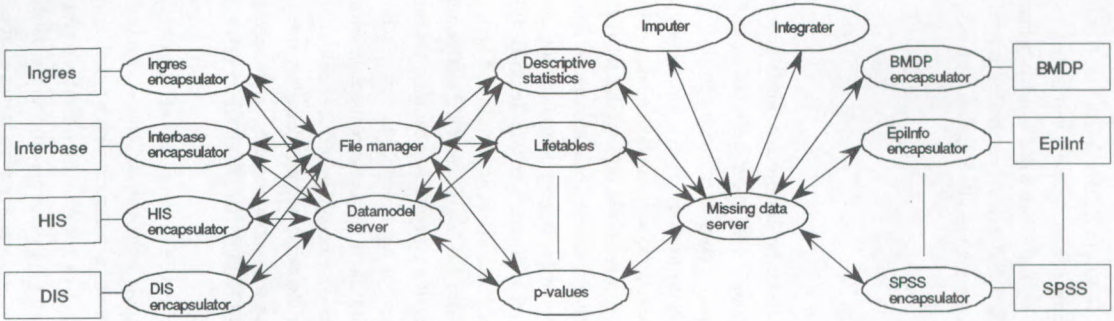
voorbeeld van Gibbs sampling, een simulatietechniek om observaties uit gecompliceerde posterior verdelingen te trekken (Gelfand en Smith, 1990). Uiteraard beperkt de normaliteitsassumptie het aantal toepassingsmogelijkheden van de Rubin-Schafer methode. Het is echter heel goed mogelijk de methode te generaliseren zodanig dat niet-lineaire verbanden, categorische data en andere afwijkingen van het lineaire model op een juiste wijze worden behandeld. Van Buuren et al. (1993) geven een voorbeeld waarin de Rubin-Schafer methode wordt uitgebreid met extra stappen die zorgen voor een optimale lineairisering van alle bivariate verbanden. Andere variaties zijn momenteel in ontwikkeling.

4 Multiple imputatie in HERMES

Het ontwerp van het HERMES werkstation komt voort uit de groeiende behoefte om bestaande softwaretoepassingen zonder modificaties in een open gedistribueerde omgeving te integreren (van Mulligen, 1993; van Mulligen, Timmers en van Bemmel, 1993a). De ruggegraat van het HERMES werkstation wordt gevormd door de combinatie van een uniforme en transparante toegang via het netwerk en de rekenkracht van hedendaagse systemen. HERMES ondersteunt het automatisch uitwisselen van gegevens en toegang tot functies in bestaande software op verschillende locaties in het netwerk middels een gebruikersvriendelijk werkstation. Uit onderzoek blijkt dat het werkstation bruikbaar is voor gebruikers die onervaren zijn in het werken met meerdere systemen (van Mulligen, Timmers en van Bemmel, 1993b).

Uit de interactie met de gebruiker stelt een client applicatie een bericht in de HERMES berichtentaal samen. Dit bericht bevat de data die de gebruiker wil analyseren, de uit te voeren functie of de te raadplegen kennis. Het bericht wordt naar een centrale 'broker service' gestuurd. Deze broker service is verantwoordelijk voor het vinden van een geschikte server om het bericht af te handelen. De broker beschikt daartoe over een speciale database die naast de defaultservices ook gebruikersvoorkeuren voor bepaalde services bevat. Nadat de broker de beste kandidaat heeft bepaald wordt het bericht aangevuld met defaultwaarden voor de verplichte parameters van de geselecteerde service en verder verstuurd. Bestaande server processen worden zoveel mogelijk hergebruikt zodat de overhead van het client-server model zo klein mogelijk wordt gehouden.

De indirecte communicatie tussen client en server maakt een flexibele aanpak mogelijk binnen continue en snel veranderende computeromgevingen. Vooral in gebieden waar gegevens snel tussen verschillende informatiesystemen getransporteerd worden, en waar voortdurend nieuwe applicaties worden geïnstalleerd (bijv. in de medische wereld) vergroot een 'late binding' tussen bericht en service de mogelijkheden gebruik te maken van de laatste ontwikkelingen. Door de functie van de broker uit de client applicatie te verplaatsen naar een centrale service wordt de hoeveelheid benodigde code voor het installeren van nieuwe services beperkt.



Figuur 2
Schema van de HERMES client-server architectuur voor multiple imputatie. Onderdelen in rechthoeken verwijzen naar commercieel verkrijgbare software. De hieraan gekoppelde encapsulators zorgen voor de vertaling tussen de HERMES berichtentaal en de specifieke applicatie commando's. De overige cellen zijn specifieke HERMES client-server modulen. De communicatie hiertussen verloopt zoals is neergelegd in de database van de broker service en verloopt volgens de aangegeven pijlen. De gebruikersmodulen 'Descriptive statistics', 'Lifetables' etc. maken slechts via de missing data server gebruik van de statistische software. De missing data server zorgt voor de imputatie-, analyse- en integratiestappen.

Bestaande applicaties kunnen worden benut via een zg. 'encapsulator'. Een dergelijk encapsulator is een vertaalprogramma dat, in beide richtingen, de HERMES berichtentaal omzet in opdrachten die de toepassing kan uitvoeren. Voor elke toepassing is een encapsulator vereist. Momenteel zijn encapsulators ontwikkeld voor het ziekenhuis informatie systeem BAZIS, een afdelingsinformatie systeem, de databasesystemen Ingres, Oracle, RDB, Interbase en dBase, en voor diverse statistische en presentatiepaketten (o.m. BMDP, SPSS, EpiInfo, WingZ en Harvard Graphics). Net als bij gewone services zijn encapsulators via het netwerk toegankelijk.

De HERMES client-server architectuur is bij uitstek geschikt voor het opnemen van een missing data server. Momenteel versturen de statistische clients hun berichten direct naar de statistische servers. Door de database van de broker aan te passen kunnen in de toekomst berichten aan de statistische servers worden onderschept en doorgezonden naar de missing data server. Op haar beurt kan de missing data server van de statistische servers gebruik maken op dezelfde wijze waarop de statistische clients dat nu doen (zie Figuur 2).

De werking van de missing data server is grofweg als volgt. De missing data server ontvangt het bericht van de client en gaat na of er ontbrekende gegevens zijn. Indien die er niet zijn wordt het bericht doorgezonden naar de statistische server en wordt het resultaat naar de client teruggezonden. Zijn er wel missing data dan bepaalt de missing data server een geschikte imputatietechniek. Dit gebeurt aan de hand van eigenschappen van de gegevens en d.m.v. interactie met de gebruiker. De werkelijke imputaties worden verzorgd door een imputer service. Voor elk van de m gecompleteerde datasets activeert de missing data server vervolgens de statistische service die door de client werd gevraagd. De resultaten worden aan de missing data server teruggezonden. Tenslotte worden alle resultaten gecombineerd tot één antwoord door de integrater service. Het eindresultaat wordt geretourneerd naar de statistische client.

5 Conclusie

Het projekt is een samenwerking tussen TNO Preventie en Gezondheid in Leiden en de Erasmus Universiteit in Rotterdam. Het is nu aangekomen in het tweede jaar van een vierjaars termijn. De belangrijkste doelen zijn gezet en tot op heden is het meeste onderzoek gedaan naar verschillende methoden voor het verkrijgen van datagestuurde imputaties. Voor categorische gegevens is een multiple versie van MISTRESS ontwikkeld (van Buuren en van Rijckevorsel, 1992). Voor niet-normaal verdeelde kwantitatieve data wordt geëxperimenteerd met regressie quantielen, hot-deck methoden en met locale aanpassingen van de Rubin-Schafer methode. Criteria voor de keuze tussen verschillende methoden zullen worden ontwikkeld op basis van uitgebreide simulatiestudies.

We denken dat de HERMES gebruikersinterface, de client-server architectuur, de faciliteiten voor bestandscombinatie en de datagestuurde aanpak van multiple imputatie een vruchtbare

combinatie vormt. Belangrijk doel van het projekt is een systeem te ontwikkelen waarin ontbrekende gegevens op een verantwoorde wijze op een routinematige manier kunnen worden behandeld. Omdat een dergelijk systeem nog niet beschikbaar is zal het systeem ook voor de statistische wereld van belang zijn. Een bijkomend doel is de HERMES omgeving te verrijken met state-of-the-art technieken, onder behoud van de voordelen van het gebruik van bestaande software. Hierdoor komen de best denkbare technieken ook binnen het bereik van statistisch minder onderlegde gebruikers. Dit is niet alleen interessant voor database specialisten, maar ook voor onderzoekers die praktijkproblemen willen oplossen met een minimum aan inspanning.

Nawoord

De oorspronkelijke Engelstalige versie van deze tekst is verschenen als S. van Buuren, E.M. van Mulligen and J.P.L. Brand (1994), Routine multiple imputation in statistical databases. In J.C. French and H. Hinterberger (Eds.), *Seventh International Working Conference on Scientific and Statistical Database Management, 28–30 Sept., Charlottesville, Virginia* (pp. 74-78). Los Alamitos: IEEE Computer Society Press. We danken professor Rubin en verscheidene anonieme reviewers voor hun commentaar op eerdere versies.

Literatuur

Gelfand, A. & Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398–409.

Little, R.J.A. & D.B. Rubin (1987). *Statistical analysis with missing data*. New York: Wiley.

Rubin D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rubin, D.B. & Schafer, J.L. (1990). Efficiently creating multiple imputations for incomplete multivariate normal data. *ASA 1990 Proceedings of the Statistical Computing Section* (pp. 83–88).

van Buuren, S. & van Rijckevorsel, J.L.A. (1992). Imputation of missing categorical data by maximizing internal consistency. *Psychometrika*, 57, 567–580.

van Buuren, S., van Rijckevorsel, J.L.A. & Rubin, D.B. (1993). Multiple imputation by splines. *Bulletin of the International Statistical Institute, Contributed Papers II*, 503–504.

van Mulligen, E.M. (1993). *An architecture for an Integrated Medical Workstation: Its realization and evaluation*. Dissertation, Dept. of Medical Informatics, Erasmus University Rotterdam. ISBN 90–9006284–X.

van Mulligen, E.M., Timmers, T. & van Bemmel J.H. (1993a). A new architecture for integration of heterogeneous software components. *Methods of Information in Medicine*, 32, 292–301.

van Mulligen, E.M., Timmers, T. & van Bemmel, J.H. (1993b). User evaluation of an integrated medical workstation for clinical data analysis. *Methods of Information in Medicine*, 32, 365–372.

Ontvangen: 11-7-1994
Geaccepteerd: 22-12-1994