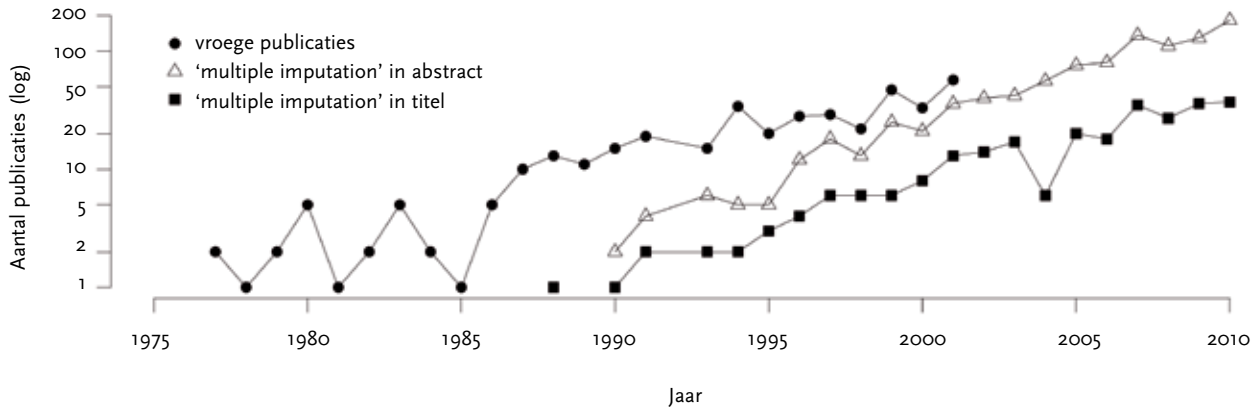
MULTIPELE IMPUTATIE IN VOGELVLUCHT

Stef van Buuren

We hebben het er liever niet over, maar allemaal worden we geplaagd door ontbrekende gegevens. Het liefst moffelen we problemen veroorzaakt door ontbrekende gegevens onder het tapijt. De standaard 'oplossing' voor missing data bestaat uit het weglaten van de onvolledige rijen uit de analyse. In veel software gebeurt dit automatisch, en als gevolg daarvan zien we vaak dat de steekproefgrootte varieert tussen verschillende tabellen, figuren en analyses. Buiten dat de rapportage minder consistent is, gaat er bij de standaard aanpak veel kostbaar verzameld materiaal verloren. Bovendien kan het weglaten van informatie tot foutieve conclusies leiden. De laatste jaren zijn betere methoden ontwikkeld voor het omgaan met incomplete gegevens. In dit artikel laat ik de lezer snuffelen aan één van deze methoden, multipele imputatie.



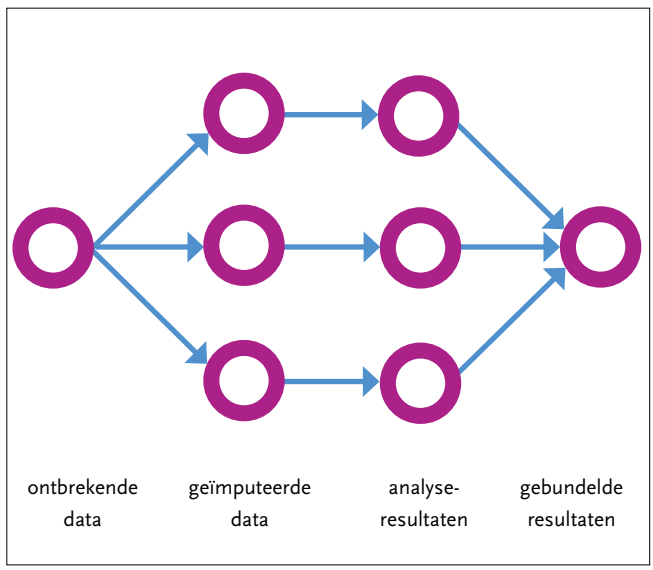
Figuur 1. Aantal publicaties (log) over multipele imputatie gedurende de periode 1977 - 2010 volgens drie telmethoden

Multipele imputatie is rond 1980 door Donald B. Rubin ontwikkeld. Figuur 1 illustreert dat multipele imputatie pas sinds kort geaccepteerd is. Multipele imputatie heeft tot doel de onzekerheid te schatten die het gevolg is van het ontbreken van informatie. Belangrijk voordeel van de methode is dat standaard analysetechnieken, die doorgaans alleen werken voor complete data, ongewijzigd kunnen worden toegepast. Onder ruime voorwaarden geeft de methode zuivere schattingen, en correcte standaardfouten, P-waarden en betrouwbaarheidsintervallen.

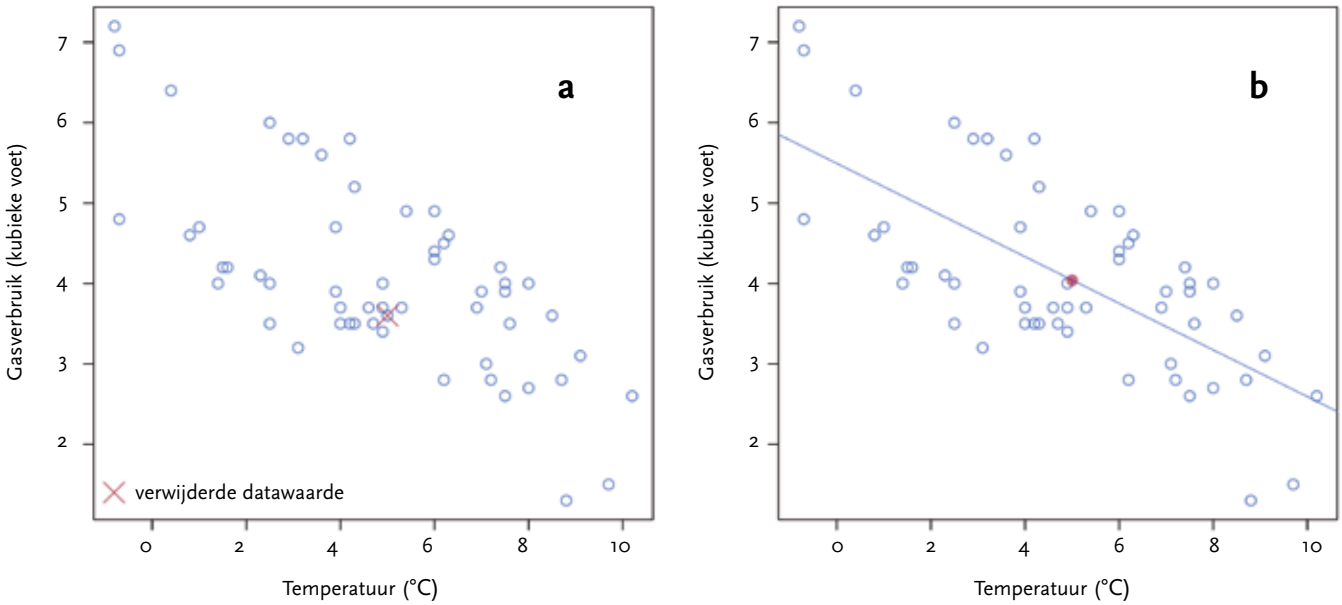
Hoe werkt multipele imputatie?

Figuur 2 beschrijft schematisch de belangrijkste stappen. De figuur start aan de linkerkant met een incomplete data set. De ontbrekende gegevens worden eerst vervangen door imputaties. Dit gebeurt driemaal, resulterend in drie verschillende versies van de compleet gemaakte dataset. Deze versies zijn identiek op de plaatsen waar de echte data staan, maar zullen verschillen onderling op de plaatsen waar de data zijn verzonnen. Hoe dit imputeren in zijn werk gaat behandelen we verderop.

Elk van de drie geïmputeerde datasets analyseren we met de methode die we hadden willen toepassen als de data compleet zouden zijn geweest. De parameterschattingen zullen onderling verschillen. Dit is lastig omdat we niet één, maar drie resultaten krijgen. Realiseer echter dat de resultaten slechts van elkaar verschillen omdat de imputaties variëren. Hoe meer gegevens ontbreken en hoe meer variatie de imputatie vertonen, hoe meer de drie resultaten onderling zullen verschillen. Deze extra variatie tellen we op bij de gebruikelijke steekproefvariatie, en hiermee kunnen we correcte betrouwbaarheidsintervallen en P-waarden berekenen.



Figuur 2. Schema van de hoofdstappen van multipele imputatie



Figuur 3a en 3b. Twee manieren om ontbrekend gasverbruik te imputeren bij 5° Celsius: a. geen imputatie, b. voorspelde waarde

Genereren van multipele imputaties

De methode om multipele imputaties te genereren moet 'proper' zijn. Het voert te ver om hier de precieze definitie van proper imputaties uit te leggen, maar in de praktijk komt proper erop neer dat het imputatiemodel rekening moet houden met het mechanisme dat de ontbrekende data creëerde, met de relaties in de gegevens, en met de onzekerheid over deze relaties.

Hoe kunnen we imputaties genereren die aan de bovenstaande criteria voldoen? We kijken hiervoor naar de *whiteside*-dataset uit het R package MASS. Whiteside noteerde gedurende twee winters (1960 en 1961) het wekelijks gasverbruik (in kubieke voet) van zijn woning in Zuid-Oost Engeland, en de gemiddelde buitentemperatuur (in Celsius).

Improper imputatie

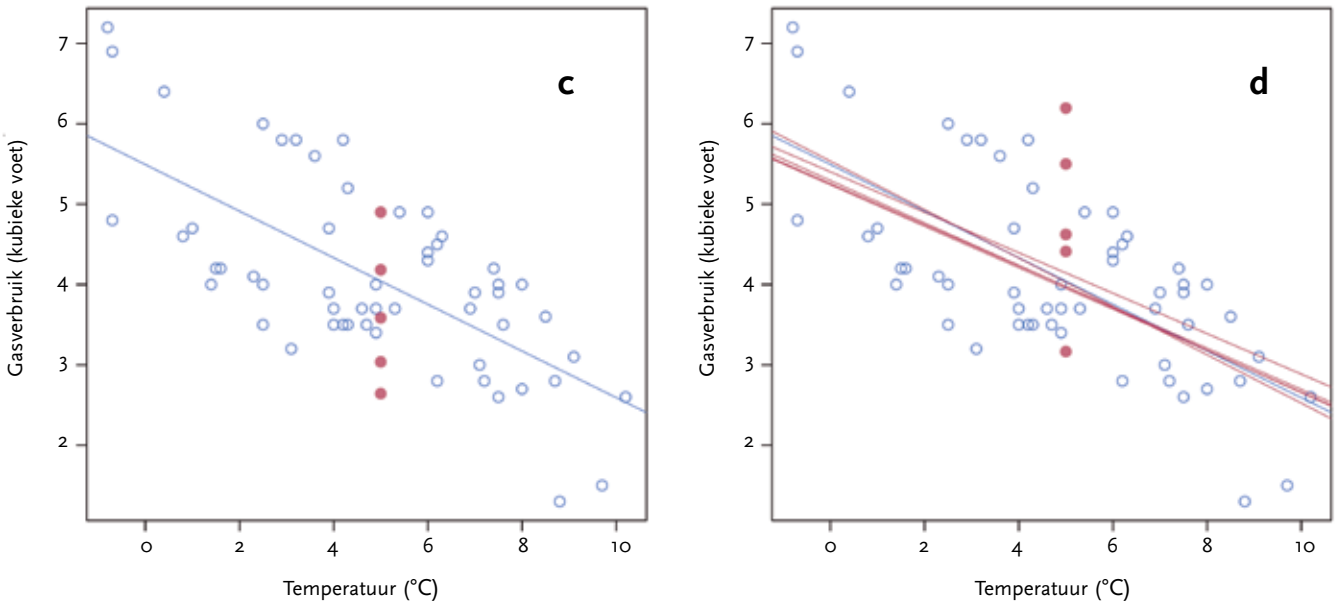
Figuur 3a is een spreidingsdiagram van de gegevens. Meer gas is nodig tijdens koudere weken, zodat er een sterk negatief verband is. De dataset is compleet, maar laten we ter illustratie de gasverbruik uit rij 47 verwijderen. De temperatuur tijdens deze week was 5° Celsius. Hoe kunnen we multipele imputaties voor het ontbrekende gasverbruik genereren?

Een eerste mogelijkheid is de regressielijn te berekenen, en de imputatie vanaf de regressielijn te nemen. De regressievergelijking is gelijk aan

$$\text{gasverbruik} = 5{,}49 + 0{,}29\ \text{temperatuur},$$

dus bij een temperatuur van 5° Celsius is de voorspelde waarde gelijk aan 5,49 - 0,29 maal 5 = 4,04. Figuur 3b laat zien waar de geïmputeerde waarde ligt. Merk op dat deze waarde de 'beste' waarde is, dat wil zeggen de meest waarschijnlijke onder het model. Echter, het is niet de beste imputatie omdat uit de waarde zelf niet blijkt wat de kwaliteit is. De voorspelde waarde geeft niet de mate van onzekerheid weer.

We kunnen de methode verbeteren door ruis toe te voegen aan de voorspelde waarde. Veronderstel



Figuur 3c en 3d. Twee manieren om ontbrekend gasverbruik te imputeren bij 5° Celsius: c. voorspelde waarde + ruis, d. voorspelde waarde + ruis + parameter onzekerheid

dat de geobserveerde gegevens normaal verdeeld zijn rond de regressielijn. De geschatte standaard deviatie is gelijk aan 0,86 kubieke voet. Het idee is een waarde willekeurig te trekken uit de normaalverdeling met gemiddelde nul en standaard deviatie 0,86, deze waarde bij de voorspelde waarde van 4,04 op te tellen, en het resultaat te gebruiken als imputatie. We kunnen het trekken uit de normaalverdeling herhalen, en daarmee multipele imputaties maken. Figuur 3c illustreert dit proces voor vijf imputaties. Gemiddeld zullen de imputaties gelijk zijn aan de voorspelde waarde. De variatie van de imputaties weerspiegelt het feit dat we gasverbruik niet exact uit de temperatuur kunnen voorspellen.

Proper imputatie

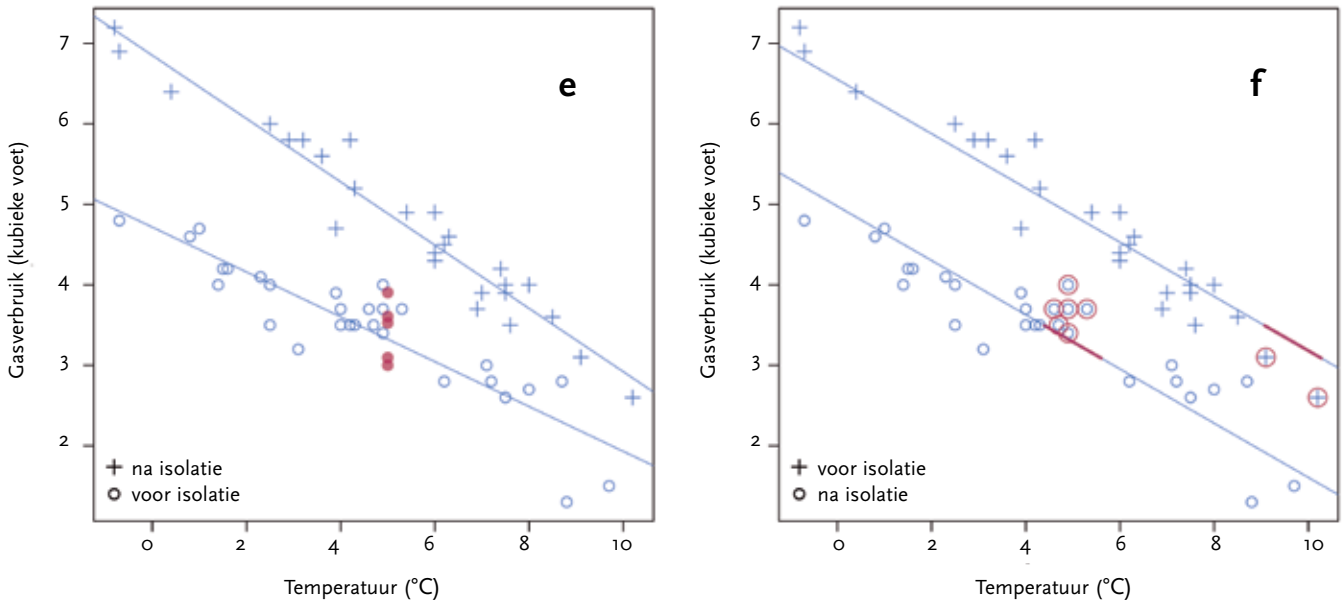
Het toevoegen van ruis is een stap voorwaarts, maar is nog niet geheel juist. De methode uit de vorige paragraaf veronderstelt dat we weten waar de regressielijn ligt. In de praktijk zijn de regressieparameters onbekend, en moeten ze geschat worden uit de data. Dat betekent echter ook dat de regressielijn zelf onderhevig is aan de steekproefvariatie.

De onzekerheid van de regressielijn kan ook in de imputaties worden opgenomen. Dat kan op twee manieren. De Bayesiaanse methode trekt de parameters uit hun posterior verdeling, gegeven de data. De bootstrap methode hertrekt eerst de observaties (met teruglegging), en schat de parameters uit deze steekproef.

Figuur 3d bevat vijf getrokken regressielijnen berekend met behulp van de Bayesiaanse methode. Een imputatie bestaat nu uit de voorspelde waarde vanaf de rode lijn, plus een trekking uit de ruisverdeling.

Scherper imputeren

De gegevens bevatten nog een tweede voorspeller die aangeeft of het huis geisoleerd was. Het opnemen van deze voorspeller in het imputatiemodel reduceert de



Figuur 3e en 3f. Twee manieren om ontbrekend gasverbruik te imputeren bij 5° Celsius: e. twee voorspellers, f. predictive mean matching

variatie van de imputaties. Figuur 3e geeft de datapunten weer met labels van de isolatiestatus. De figuur bevat twee regressielijnen, eentje (de bovenste) voor het gasverbruik van het niet-geïsoleerd huis, de andere voor het verbruik na isolatie. Na isolatie is het gasverbruik aanzienlijk lager. Stel dat we ook weten dat het huis geïsoleerd is. Hoe zouden we dan de imputatie moeten trekken?

We passen dezelfde methode toe als hierboven, maar nu gebruik makend van de onderste regressielijn. Figuur 3e laat de vijf imputaties zien van deze methode. Zoals verwacht is de verdeling van de imputaties gemiddeld lager. Merk op dat de variatie tussen de imputaties nu kleiner is. We kunnen het gasverbruik nauwkeuriger inschatten, en de onzekerheid wordt daarmee minder.

Figuur 3f, tenslotte, illustreert een alternatief. Gelijk aan de eerdere methode berekenen we het voorspelde gasverbruik bij 5° Celsius voor het geïsoleerde huis. We selecteren vervolgens een klein aantal kandidaat donoren (meestal 3 of 10). Deze donoren worden zodanig gekozen dat zij hun voorspelde waarde gelijk is aan, of dichtbij ligt, bij de voorspelde waarde van het te imputeren record. Uit deze kandidaat donoren trekken we willekeurig een donor. We gebruiken het geobserveerde gasverbruik van deze donor als imputatie. Deze methode staat bekend als predictive mean matching. Een prettige eigenschap is dat zij altijd imputaties levert van waarden die werkelijk geobserveerd zijn. Deze methode is simpel en bijzonder robuust tegen schendingen van de lineariteitsassumptie van het imputatiemodel.

Multivariate missing data

In de praktijk komen de ontbrekende gegevens in meerdere variabelen voor. Hoe kunnen we dan imputaties genereren? Multivariate Imputation by

Chained Equations (MICE) biedt hiervoor een eenvoudige oplossing. Stel dat we als start elke ontbrekende waarde invullen met een willekeurige trekking uit de geobserveerde data. Het MICE-algoritme imputeert de eerste variabele in de dataset zoals boven beschreven, onder de tijdelijke aanname dat alle andere variabelen compleet zijn, imputeert dan de tweede variabele gebruikmakende van de eerdere imputaties op de eerste variabele, etc. Na 5 iteraties is het algoritme dikwijls geconvergeerd. Deze eenvoudige methode is flexibel en produceert imputaties van hoge kwaliteit.

Software en documentatie

Het boek van Rubin (1987) geeft de statistische onderbouwing van de techniek. Het MICE package in R (Van Buuren en Groothuis-Oudshoorn, 2011) bevat de meest uitgebreide implementatie van het MICE algoritme. Vereenvoudigde versies van het algoritme zijn de laatste twee jaar beschikbaar gekomen in IBM SPSS, SAS en Stata. Mijn onlangs verschenen boek *Flexible Imputation of Missing Data* (2012) beschrijft de techniek en de methodologie in detail. In combinatie met het MICE package kan de geïnteresseerde lezer daarmee meteen aan de slag.

Literatuur

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Van Buuren, S. & Groothuis-Oudshoorn, C. G. M. (2011). mice: Multivariate Imputation by Chained Equations. *Journal of Statistical Software*, 45(3), 1–67.

Van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Boca Raton, Florida: Chapman and Hall/CRC Press.

Stef van Buuren is bijzonder hoogleraar Applied Statistics in Prevention, verbonden aan de Faculteit Sociale Wetenschappen, Afdeling Methoden en Statistiek van de Universiteit Utrecht, en senior onderzoeker bij TNO waar hij hoofd van de afdeling statistiek is.
E-mail: <stef.vanbuuren@tno.nl>, <S.vanBuuren@uu.nl>