

Evaluatie en verbetering van de datakwaliteit

Kwantitatieve beslissingsvorming bij grote financiële instelling in België

Er is sprake van een trend naar meer kwantitatieve verwerking van data. Het resultaat van een kwantitatieve analyse is echter maar zo goed als de data die men invoert en ontoereikende datakwaliteit kan tot hoge kosten leiden. Dit geldt bijvoorbeeld voor de financiële sector, waar beslissingen sterk afhankelijk zijn van de kwaliteit van de aangeleverde data, zeker als het gaat om de inschatting van de kredietwaardigheid van een tegenpartij.

Karel Dejaeger, Jessica Ruelens, Tony Van Gestel,

Joachim Jacobs, Bart Baesens, Jonas Poelmans en Bart Hamers

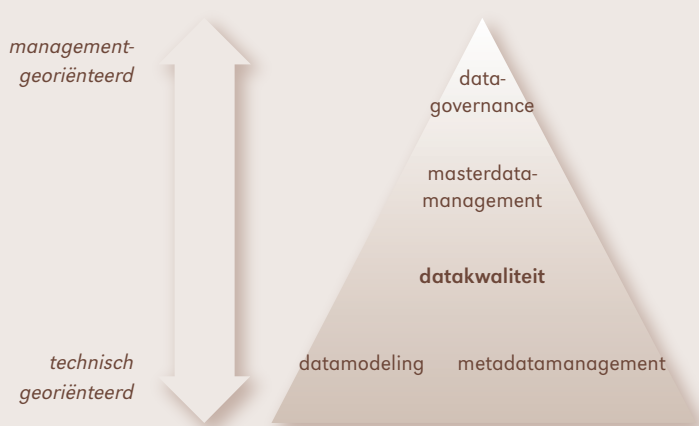
Data vormen de basis van waaruit kennis wordt opgebouwd en zijn dus een essentiële factor in het welvaren van een bedrijf. In de bedrijfsweld is er een trend naar meer kwantitatieve verwerking van de data. Deze trend wordt mede ondersteund door overheidsregulering zoals Basel II (voor financiële instellingen) en Insurance Solvency I (voor verzekeringen). Deze afspraken leggen onder meer vast dat instellingen op een bepaalde wijze gegevens moeten verzamelen en analyseren. Ook in andere takken van de bedrijfsweld worden gegevens steeds meer op kwantitatieve wijze geanalyseerd, zoals in de marketing (denk hierbij aan cross-sellinganalyses, customer segmentation en churn management) en de auditing (het opsporen van fraude door te steunen op Benford's Law). Zelfs op plaatsen waar men het niet onmiddellijk verwacht, worden data op kwantitatieve wijze verwerkt, zoals in ziekenhuizen. Dergelijke systemen worden aangeduid met de term 'clinical decision support systems' en hoewel deze zeker nog niet helemaal

tot in de laatste functionele details in orde zijn, is het de bedoeling dat op basis van ingevoerde gegevens ondersteunende informatie naar de arts toe wordt gegenereerd. Uiteindelijk zou deze informatie de arts moeten kunnen bijstaan bij het nemen van belangrijke beslissingen (Berlin, Sorani & Sim, 2006).

Het resultaat van een kwantitatieve analyse is echter maar zo goed als de data die men invoert. Vaak zullen deze data van lagere kwaliteit zijn, wat een niet te onderschatten probleem vormt. Bedrijven hebben de neiging de kosten van ontoereikende datakwaliteit te onderschatten. Deze kosten kunnen zich op verschillende manieren manifesteren en het blijkt dan ook niet eenvoudig om de precieze kosten van ontoereikende datakwaliteit te berekenen. Een typisch voorbeeld hiervan zijn personeelskosten. Voordat men data kan gebruiken in het bedrijf, zal men vaak aandacht moeten besteden aan de kwaliteit van de data. Men zal onder andere ontbrekende

Samenvatting

Datakwaliteit is van groot belang voor alle gebruikers en wordt bij voorkeur geanalyseerd op verschillende dimensies (kwaliteitsassen). Datagovernance kan hierin een cruciale rol spelen, ondersteund door technieken als metadatamanagement, masterdatamanagement en datamodeling. Datakwaliteit is een voortdurende opgave die vanaf de eerste stap van het proces tot het eindgebruik van de data prioriteit zou moeten krijgen.



Figuur 1. De verschillende aspecten binnen databeheer

velden moeten aanvullen en dubbele gegevens eruit moeten filteren. Ook kan lage datakwaliteit ertoe leiden dat de dataverwerking vertraging oploopt en zelfs opnieuw moet worden opgestart. Enkele studies hebben getracht de verschillende kostenaspecten van datakwaliteit op kwantitatieve wijze te identificeren en kwamen tot sprekende cijfers (Eckerson, 2002; PA Consulting, 2007). De kosten van slechte datakwaliteit binnen de banksector worden geraamd op 25 miljard dollar in 2008 (Kopp, 2006). De banksector vereist een nog hogere datakwaliteit dan de meeste andere sectoren. Dit komt tot uiting in onder meer de Basel II-regeling, die van financiële instellingen volledige transparantie en traceerbaarheid van data vraagt. Het nemen van beslissingen in de financiële sector is dan ook sterk afhankelijk van de kwaliteit van de aangeleverde data, zeker als het gaat om de inschatting van de kredietwaardigheid van een tegenpartij. Het opstellen en jaarlijks controleren van de hiervoor gebruikte modellen wordt opgelegd door de Basel II-regeling (Basel Committee on Banking Supervision, 2006).

Datamanagement

Om hoge kwaliteitsdata te verkrijgen is een goed databeheer belangrijk. Het domein van databe-

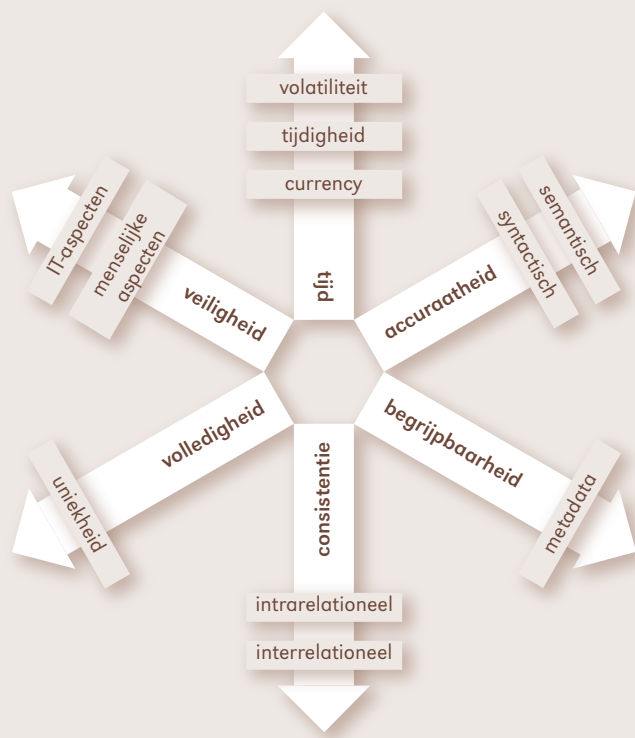
heer omvat diverse deelaspecten waarvan sommige een meer technische en andere een meer managementgeoriënteerde inslag hebben (zie figuur 1). De sleutel tot een zorgvuldig databeheer ligt in het opzetten van een datagovernancestructuur binnen de organisatie. Databeheer heeft tevens behoefte aan meer technische hulpmiddelen zoals masterdatamanagement, metadatamanagement en datamodeling.

Datagovernance wordt door het Data Governance Institute gedefinieerd als: 'Data Governance is the exercise of decision-making and authority for data-related matters' (Thomas, 2008). Datagovernance is dus een continu programma om de data binnen een organisatie te beheren en de datakwaliteit te optimaliseren. Communicatie is hierbij essentieel en behelst zowat 80 tot 95 procent van de totale inspanning die het project vereist (Hopwood, 2008). Tevens belangrijk is het besef dat dit programma niet enkel weggelegd is voor IT. Integendeel, de business moet er actief bij betrokken worden omdat hun kennis van de manier van werken tot belangrijke inzichten kan leiden. Historisch gezien hanteren de meeste financiële instellingen wel één of ander soort van datagovernanceprogramma, maar binnen de meeste financiële instellingen is dit nog niet formeel uitgebouwd. Men moet de overgang maken naar een datacultuur en datacoördinatie door de hele bank heen; de ondersteuning vanuit het management is hierbij cruciaal.

Masterdatamanagement is het beheer van de data die belangrijke entiteiten binnen het bedrijf beschrijven. Bedrijven worden opgedeeld in departementen die elk eigen applicaties, methoden en procedures gebruiken. Door deze opsplitsing wordt eenzelfde concept op verschillende manieren voorgesteld binnen de afzonderlijke applicaties. Dit worden 'islands of information coherence' genoemd (Loshin, 2008). Het gevolg is dat er meerdere, vaak disparate dataverzamelingen zijn die hetzelfde concept moeten voorstellen. De oplossing voor dit probleem is het definiëren



van een consistente en uniforme set attributen die belangrijke entiteiten binnen een bedrijf beschrijven (White e.a., 2006). Het eindresultaat van deze inspanning zal de vorm krijgen van een 'woordenboek' dat aanvaard wordt door het hele bedrijf (ook wel Shared Business Vocabulary, SBV genoemd). De gegevens uit de verschillende systemen kunnen dan gestandaardiseerd worden volgens de opgestelde SBV. De resulterende verzameling data-elementen wordt de masterdata genoemd. *Metadatamanagement* is het beheer van gestructureerde informatie die informatiebronnen beschrijft en lokaliseert of het gemakkelijker maakt om informatiebronnen op te halen, te gebruiken en te beheren (NISO, 2004). Er zijn verschillende soorten metadata denkbaar. Vaak wordt er een onderscheid gemaakt tussen drie soorten (Kim, 2005). Een eerste vorm van metadata is technische metadata, waarmee technische gegevens over de data bedoeld worden, zoals in welk dataformaat de metadata opgeslagen zijn en welke compressietechnieken er gebruikt zijn. Een ander metadata-type is operationele metadata; hiermee doelt men op de lifecycledata van de opgeslagen data, zoals gegevens over creatie en wijzigingen van data. Een laatste vorm van metadata is businessmetadata, die de beschrijving van de opgeslagen data zelf (wie, wat, waar et cetera) behelst. Bij *datamodeling* wordt er een logisch model van de (vereenvoudigde) realiteit gecreëerd. Meestal wordt een dergelijk model op papier gezet aan de hand van een EER (Enhanced Entity Relationship)- of UML (Unified Modeling Language)-diagram en vormt het de basis waarop het databasesysteem wordt gebouwd. Als dusdanig zal een goed datamodel impact hebben op de kwaliteit van de data (Hay, 2003). In de meeste situaties echter wordt men geconfronteerd met reeds bestaande systemen en zal het onmogelijk blijken om de structuur van het datamodel te wijzigen. Een databeheerproject zal gericht zijn op datakwaliteit en zal aan de hierboven beschreven aspecten aandacht geven. Datakwaliteit op zich is echter een zeer breed begrip. Data moeten onder meer begrijpbaar, volledig, tijdig en accuraat zijn. We stellen op basis van de literatuur (Baesens, 2007; Batini & Scannapieco, 2006) een origineel



Figuur 2. Het datakwaliteitsassenkruis

datakwaliteitsassenkruis voor dat de verschillende dimensies weergeeft waaraan datakwaliteit kan worden getoetst (zie figuur 2).

We bespreken kort de verschillende aspecten van datakwaliteit. Datakwaliteit kan voorgesteld worden als bestaande uit zes verschillende deelaspecten: tijd, accuraatheid, begrijpbaarheid, consistentie, volledigheid en veiligheid.

Tijd

De dimensie tijd betreft de tijdsgelateerde aspecten van datakwaliteit en bestaat uit de volgende deelaspecten:

- *Currency* behelst de vraag of de data ogenblikkelijk worden geüpdatet indien er een verandering optreedt in het real-life fenomeen. De currency van een papieren bron zal lager zijn dan die van een elektronische bron die continu wordt vernieuwd.
- *Volatiliteit* beschrijft hoe frequent data variëren in de tijd. Stabiele data zoals geboortedatum hebben een volatiliteit van nul. Beurscijfers hebben een hoge mate van volatiliteit omdat ze enkel voor korte tijdsperiodes gelden.
- *Tijdigheid* beschrijft hoe actueel de data zijn voor de taak die men ermee moet uitvoeren. Deze dimensie wordt gemotiveerd door het feit dat het mogelijk is om actuele data te hebben die toch nutteloos zijn omdat het te laat is voor een specifiek gebruik.

Accuraatheid

Accuraatheid is de mate van correctheid van een voorstelling v' van een real-life fenomeen v . Binnen deze dimensie kan onderscheid worden gemaakt tussen syntactische en semantische accuraatheid. Het definitiedomein D is de verzameling van alle syntactisch correcte elementen om real-life fenomenen te beschrijven.

- *Syntactische accuraatheid* is de nabijheid van een waarde v' tot de elementen van het corresponderende definitiedomein D . Zo is een kredietrating AAA syntactisch correct en een rating ABC syntactisch incorrect.
- *Semantische accuraatheid* is de nabijheid van een waarde v' tot de echte waarde v . Een rating AAA is dan wel syntactisch correct, in het geval van een niet-kredietwaardige tegenpartij zal een AAA-rating niet semantisch correct zijn.

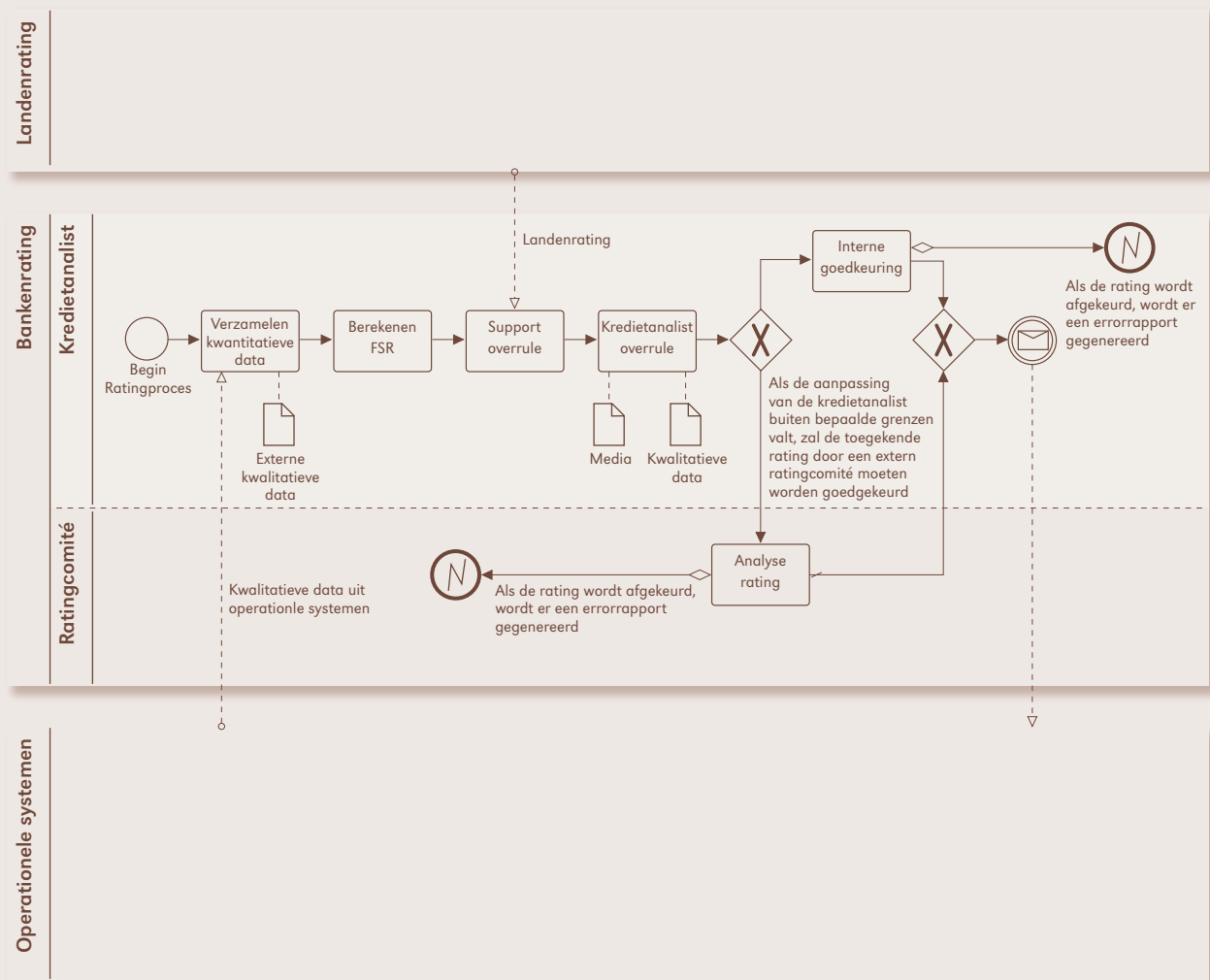
Begrijpbaarheid

Met begrijpbaarheid wordt bedoeld of de eindgebruiker de data kan begrijpen. Een goede begrijpbaarheid van de data wordt verkregen door uitgebreide datadefinities in de metadata, conformiteit met gestandaardiseerde data-uitwisselingsformaten en een heldere en consistente syntaxis.

Consistentie

Een dataset is consistent als de verbanden (*constraints*) binnen een tuple worden nageleefd. Hier kan onderscheid worden gemaakt tussen interrelationele en intrarelationele constraints.

- *Interrelationele consistentie* betreft het naleven van regels over meerdere records heen. Hiertoe moet men regels definiëren waaraan records over de dataset heen moeten voldoen.
- *Intrarelationele consistentie* verifieert of regels die van toepassing zijn binnen één record



Figuur 3. Mogelijke procesvoorstelling van het inschatten van de kredietwaardigheid van een banktegenpartij (BPMN)



gerespecteerd worden. Zo kan men bijvoorbeeld onderzoeken of een variabele binnen correcte grenzen valt.

Volledigheid

Volledigheid kan worden gedefinieerd als de mate waarin er geen ontbrekende waarden (*missing values*) zijn. Een ontbrekende waarde kan causaal of niet causaal zijn. Een causale missing value is toegestaan en betekent dat er een specifieke geaccepteerde reden is voor het ontbreken van de waarde. Het veld van het btw-nummer in een dataset zal bijvoorbeeld leeg zijn als het een particuliere persoon betreft.

Uniekheid kan worden gezien als een aanvulling van volledigheid en gaat na of er geen dubbele waarden in een dataset zitten.

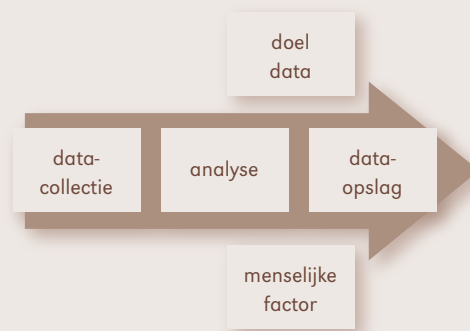
Veiligheid

Veiligheid is van groot belang voor financiële instellingen. Binnen het veiligheidsaspect kan er onderscheid worden gemaakt tussen IT-elementen (zoals wachtwoorden) en menselijke elementen (zoals functiescheiding).

Methodologie

Bij aanvang van de analyse van de datakwaliteit binnen een proces is het van belang de datastromen en communicatiekanalen in kaart te brengen: een gestroomlijnde, vlotte communicatie vormt immers een essentieel onderdeel van een goed databeheer. Hiertoe kan men de processen op een grafisch formele wijze voorstellen in BPMN (Business Process Modeling Notation) (De Backer & De Backer, 2008). De verschillende deelnemende partijen worden voorgesteld door horizontale banden en de communicatie tussen de partijen door stippellijnen en specifieke symbolen. In figuur 3 wordt weergegeven hoe binnen een financiële instelling de kredietwaardigheid van een bank zou kunnen worden bepaald.

Analyses van processen zijn steeds onvolledig zonder diepgaande contacten met de betrokkenen op de werkvloer. Interviewsessies met deze personen zijn van onschatbare waarde om het proces in kaart te brengen. Om bovendien eventuele problemen met de datakwaliteit te identificeren is het wense-



Figuur 4. De verschillende dimensies van de vragenlijst

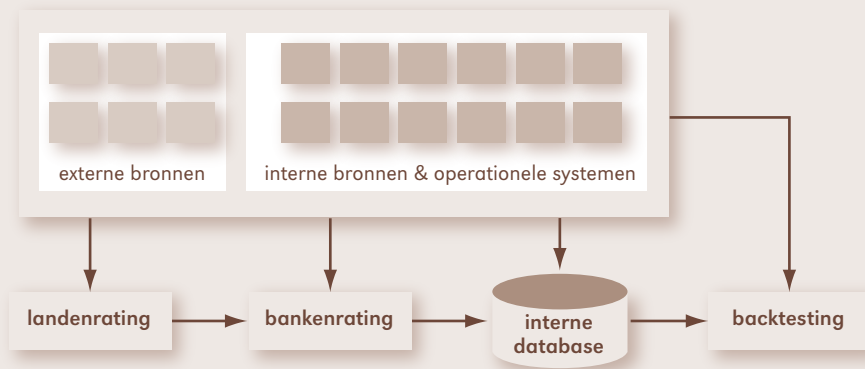
lijk deze interviews af te nemen aan de hand van een vragenlijst met een specifieke focus op datakwaliteit. We stellen een originele vragenlijst voor bestaande uit 65 vragen. Deze vragenlijst is opgebouwd vanuit het idee dat data binnen welk proces dan ook eerst worden verzameld. Hierna zullen de data worden verwerkt of geanalyseerd, om in een derde stap te worden opgeslagen en doorgestuurd naar een volgende schakel in de keten (AHIMA, 1998) (zie figuur 4). Aan de vragenlijst werden verder nog een menselijke dimensie (opleiding en communicatie) en een doeldimensie (is het doel van de dataverwerking duidelijk?) toegevoegd. Op basis van de BPMN-procesvoorstelling is het mogelijk de vragen beter te situeren. De vragen werden over de verschillende dimensies heen gestructureerd; binnen deze dimensies zijn de vragen opgedeeld volgens de definitie van datakwaliteit (cfr. het datakwaliteitsassenkruis). Afhankelijk van de dimensie worden andere datakwaliteitsaspecten belangrijker geacht. Het is mogelijk om aan de vragen een score toe te kennen. Door middel van deze score kunnen de resultaten dan per dimensie met beschrijvende statistieken zoals radarplots worden voorgesteld. (De originele vragenlijst is bij de auteurs aan te vragen.)

Resultaten

Beschrijving van proces

Een BPMN-procesvoorstelling in combinatie met een vragenlijst specifiek gericht op de evaluatie van datakwaliteit kan worden toegepast op zeer uiteenlopende processen. Concreet werd deze aanpak hier aangewend voor de evaluatie van een ratingproces binnen een grote financiële instelling in België (zie figuur 5).

Elke financiële instelling zal aan zakenrelaties (zoals andere banken) een kredietwaardigheidsrating verlenen. Om aan een bank een betrouw-

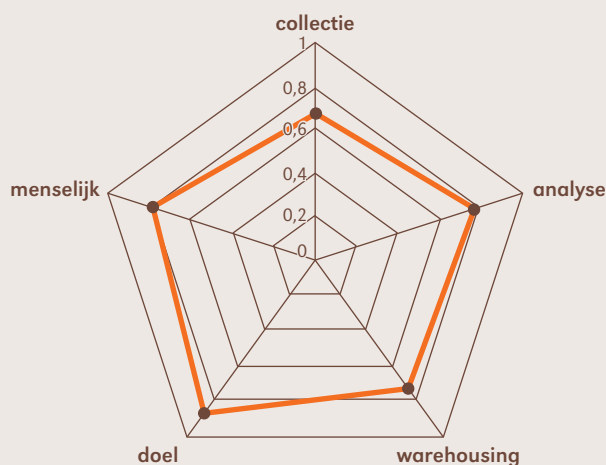


Figuur 5. High-level voorstelling van het ratingproces voor banken en van het backtestingproces

bare rating te kunnen toekennen zal ook het land waarin de bank actief is een belangrijke factor zijn (onder meer voor het berekenen van de ‘country support’). Bij het vastleggen van zowel een landenrating als een bankenrating is bijkomende interne en externe informatie van essentieel belang. De toegekende ratings worden opgeslagen in een database die ter beschikking staat van de eindgebruiker. Bij beide ratingprocessen wordt gebruikgemaakt van wiskundige modellen. Deze modellen moeten volgens de Basel II-regeling jaarlijks worden gecontroleerd (hiernaar wordt verwezen als ‘backtesting’). Voor deze controle zijn diverse elementen vereist, waaronder de toegekende ratings, de oorspronkelijke input voor de ratingmodellen en het aantal tegenpartijen dat in default (wanbetaling) is gegaan.

Semikwantitatieve voorstelling

De vragenlijst laat een semikwantitatieve aanpak toe waardoor op een meer systematische wijze foutgevoelige elementen kunnen worden gedetecteerd. In figuur 6 wordt een fictief voorbeeld

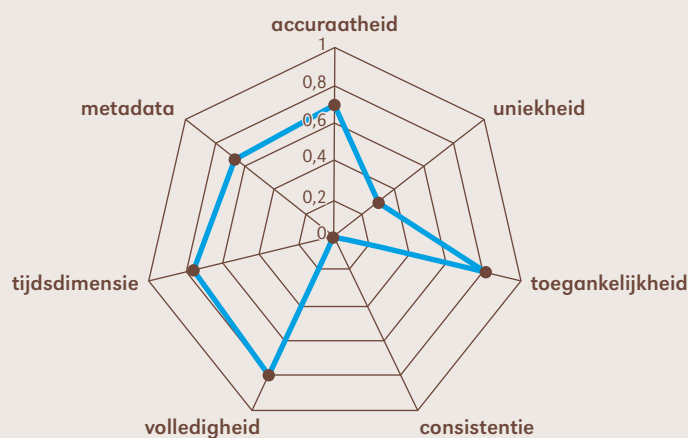


Figuur 6. Radarplot van de vijf dimensies

van een radarplot getoond met betrekking tot de vijf dimensies van de vragenlijst, waarbij de scores geaggregeerd werden per dimensie. Een lagere score duidt op een potentieel probleemgebied. In de vragenlijst kunnen ook radarplots van de afzonderlijke dimensies gegenereerd worden. Deze bevatten meer details over de verschillende aspecten van de dimensie. De dimensie ‘collectie’ bestaat uit zeven deelaspecten (zie figuur 7). De nulscore voor consistentie is te wijten aan het niet invullen van de betreffende vragen. In dit geval zou er bijvoorbeeld een probleem kunnen zijn met het voorkomen van dubbele data in de ontvangen dataset in deze stap van het proces (vergelijk de lage score op ‘uniekheid’).

De dimensie ‘analyse’ bestaat uit drie deelaspecten (zie figuur 8). Dergelijk hoge scores duiden erop dat er geen grote problemen zijn in deze fase.

De dimensie ‘warehousing’ is opgebouwd uit vijf deelaspecten (zie figuur 9). Het aspect ‘veiligheid’ mag in het geval van financiële instellingen niet verwaarloosd worden. In dit voorbeeld werden de vragen over tijdigheid niet beantwoord.



Figuur 7. Radarplot van de dimensie ‘collectie’



Bespreking

Opvallend bij de analyse is de discrepantie tussen de problemen die op het eind van het proces ervaren worden (bij de backtesting) en de resultaten van de analyse. Over het algemeen bleken de resultaten van de afzonderlijke deelprocessen goed en dit uitte zich in hoge scores in de radarplots. Hierdoor is het niet zo evident om duidelijk de oorzaken van de problemen die bij het backtesten van de modellen worden ervaren te onderkennen. Er zijn echter enkele (soms) kleinere aspecten die voor verbetering vatbaar zijn. Het cumulatief effect van deze elementen wordt vaak onderschat, waardoor de eindgebruikers van de gegevens vaak geconfronteerd worden met data van onvoldoende kwaliteit.

Een essentiële input voor het backtesten van modellen is een defaultlijst. Deze bevat gegevens over de tegenpartijen die in default gegaan zijn (de wanbetalers). Bepalen of een tegenpartij in default gaat is een commerciële aangelegenheid: men kan er bijvoorbeeld voor kiezen om de klant bijkomend betalingsuitstel te geven. Soms is dit echter geen afdoende oplossing. Hierdoor kunnen in-defaultgegevens herzien worden nadat al met het backtesten begonnen is. Er treedt dus een timingprobleem op.

Bij het toekennen van ratings is het vaak zo dat bepaalde input nog handmatig dient te gebeuren. Dit is zeker zo als de kredietanalist kwalitatieve

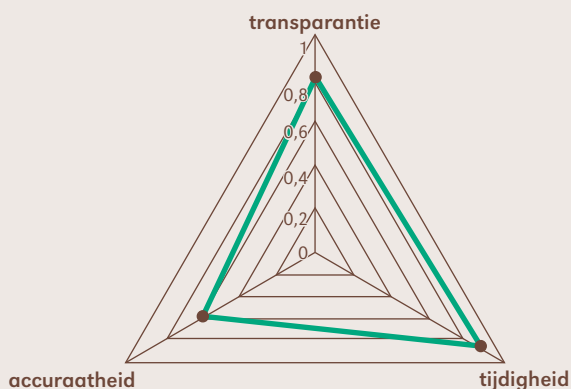
factoren wenst mee te nemen in de rating die wordt toegekend door het model. Binnen de backtesting wordt ook gebruikgemaakt van handmatig door de kredietanalist bijgehouden bestanden. Het handmatig invullen van gegevens zou zoveel mogelijk vermeden moeten worden omdat dit data van lagere accuraatheid tot gevolg kan hebben. Op het vlak van begrijpbaarheid kunnen zich problemen voordoen met de leesbaarheid van de eindverslagen. In deze nota's wordt frequent vakjargon gebruikt. Zonder een specifieke opleiding lijken deze verslagen die de motivatie zijn voor het toekennen van een bepaalde rating, niet altijd goed begrijpbaar.

De invoer van de data in bepaalde statistische modellen gebeurt soms zonder dat deze gevalideerd wordt tegen toepasselijke invoerregels. Hierdoor kan men niet zeker zijn dat de consistentieregels die van toepassing zijn op de dataset, ook worden nageleefd.

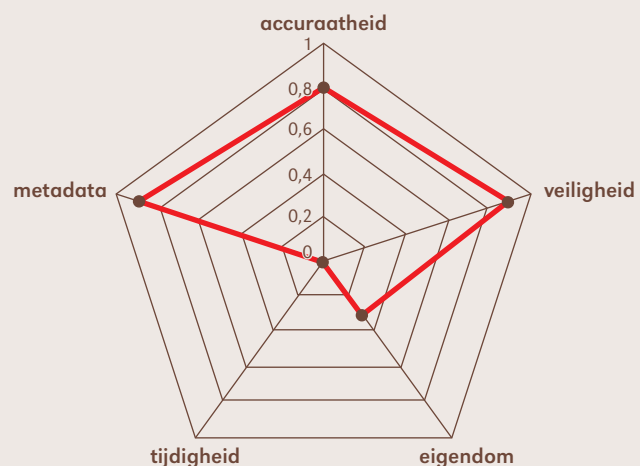
In de centrale database blijkt dat bepaalde tegenpartijen meerdere keren kunnen voorkomen. Dit kan verstrekende gevolgen hebben, zoals het tweemaal opzijzetten van risicokapitaal op dezelfde uitstaande lening. Dit zal een negatieve impact hebben op de winstgevendheid van de financiële instelling. Dergelijke problemen vallen onder het aspect 'uniekheid' van de dimensie 'volledigheid'. Wat betreft de dimensie 'veiligheid' bleken er zich geen problemen voor te doen binnen het geanalyseerde proces.

Bemerkingen

Tijdens het modelleren van de verschillende processen en de datastromen hiertussen bleek dat de Business Process Modeling Notation (BPMN) op sommige vlakken tekortschiet. Het voorstellen



Figuur 8. Radarplot van de dimensie 'analyse'



Figuur 9. Radarplot van de dimensie 'warehousing'

»Datakwaliteit is een voortdurende opgave die vanaf de eerste stap van het proces tot het eindgebruik van de data prioriteit zou moeten krijgen«

van de verschillende datastromen was niet evident omdat in BPMN de nadruk ligt op de modellering en optimalisatie van processen en minder op de datastromen.

Tevens pleiten we ervoor dat de interviews worden afgenomen door een persoon met voldoende interviewervaring. Personen zullen vaak de neiging vertonen om eigen tekortkomingen onder te belichten. Dit kwam ook tot uiting bij de analyse van de resultaten. Het eigen proces werd steeds als (bijna) ideaal voorgesteld. De focus van de personen is vaak beperkt tot hun eigen vakgebied. Dit wordt ook wel 'silodenken' genoemd. Om dit tegen te gaan is het nodig de processen te onderzoeken en de datastromen in kaart te brengen. Zo zal men kunnen streven naar een meer globaal datagebruik.

Besluit

Datakwaliteit is onmiskenbaar van groot belang voor alle gebruikers en wordt bij voorkeur geanalyseerd op verschillende dimensies (kwaliteitsassen). Uit literatuur blijkt dat datagovernance hierin een cruciale rol kan spelen en dit dient ondersteund te worden door bepaalde technieken zoals metadatamanagement en masterdatamanagement. Datakwaliteit is een voortdurende opgave die vanaf de eerste stap van het proces tot het eindgebruik van de data prioriteit zou moeten krijgen. Bij elke stap dient de nodige aandacht, kennis en motivatie aanwezig te zijn, dit gesteund door de noodzakelijke gestandaardiseerde procedures, die ook praktisch haalbaar moeten zijn en gedragen worden door de gehele organisatie. Een grondige analyse van het ratingproces voor banken heeft aangetoond dat de algemene werkwijze binnen dit proces zeker aanvaardbaar tot goed verloopt, maar toch blijken bepaalde aspecten nog voor optimalisatie in aanmerking te komen. Elke oplossing dient op maat voor het specifieke probleem ontworpen te worden. Deze oplossing kan bestaan uit technische elementen (data profiling en data cleansing), maar het menselijke aspect mag niet uit het oog verloren worden. Cruciale elementen hierbij zijn communicatie en betrokkenheid van het management.

Reviewer **Bart Baesens**

Karel Dejaeger

is assistent beleidsinformatica aan de K.U.Leuven. E-mail: karel.dejaeger@econ.kuleuven.be.

Jessica Ruelens

is afgestudeerd als handelsingenieur in de beleidsinformatica aan de K.U.Leuven. E-mail: jessicaruelens@hotmail.com.

Tony Van Gestel

is director Basel II bij Dexia. E-mail: tony.vangestel@dexia.com.

Joachim Jacobs

is quantitative analyst bij Dexia. E-mail: joachim.jacobs@dexia.com.

Bart Baesens

is docent beleidsinformatica aan de K.U.Leuven. E-mail: bart.baesens@econ.kuleuven.ac.be.

Jonas Poelmans

is assistent beleidsinformatica aan de K.U.Leuven. E-mail: jonas.poelmans@econ.kuleuven.ac.be.

Bart Hamers

is senior quantitative analyst bij Dexia. E-mail: bart.hamers@dexia.com.

Literatuur

- AHIMA. (1998). *Practice Brief: A Checklist to Assess Data Quality Management Efforts*.
- Baesens, B. (2008). It's the data, you stupid. *Data News* nr. 19, 16 april 2008.
- Basel Committee on Banking Supervision. (2006). *International convergence of capital measurement and capital standards*. Basel: Bank for International Settlements.
- Batini, C. & M. Scannapieco (2006). *Data Quality: concepts, methodologies and techniques*. Springer.
- Berlin, A., M. Sorani & I. Sim (2006). A taxonomic description of computer-based clinical decision support systems. *Journal of Biomedical Informatics* 39, pp. 656-667.
- De Backer, M. & C. De Backer (2008). Inleiding in BPMN: karakteristieken en mogelijkheden van de modelleringstaal. *Informatie* 51/3 (april), pp. 32-38.
- Eckerson, W. (2002). *Data Warehousing Special Report: Data quality and the bottom line*. Data Warehouse Institute.
- Hay, D. (2003). Data Model Quality: Where Good Data Begins. *Cutter IT Journal* vol. 16, no. 1, January 2005.
- Hopwood, P. (2008). Datagovernance: One Size Does Not Fit All. *Information Management Magazine*, June.
- Kim, W. (2005). On Metadata Management Technology: Status and Issues. *Journal of Object Technology*, vol. 4, no. 2, March-April 2005, pp. 41-47.
- Kopp, G. (2006). *Data Governance: Banks Bid for Organic Growth*. TowerGroup.
- Loshin, D. (2008). *Master Data and Master Data Management: an introduction* (white paper). DataFlux.
- NISO. (2004). *Understanding Metadata*. NISO Press.
- PA Consulting. (2007). Poor Data Quality cost 100 largest Danish companies 4 Billion DKK. Denemarken.
- Thomas, G. (2008). *The DGI Data Governance Framework*. Data Governance Institute.
- White, A. e.a. (2006). *Mastering Master Data Management*. Gartner Research.