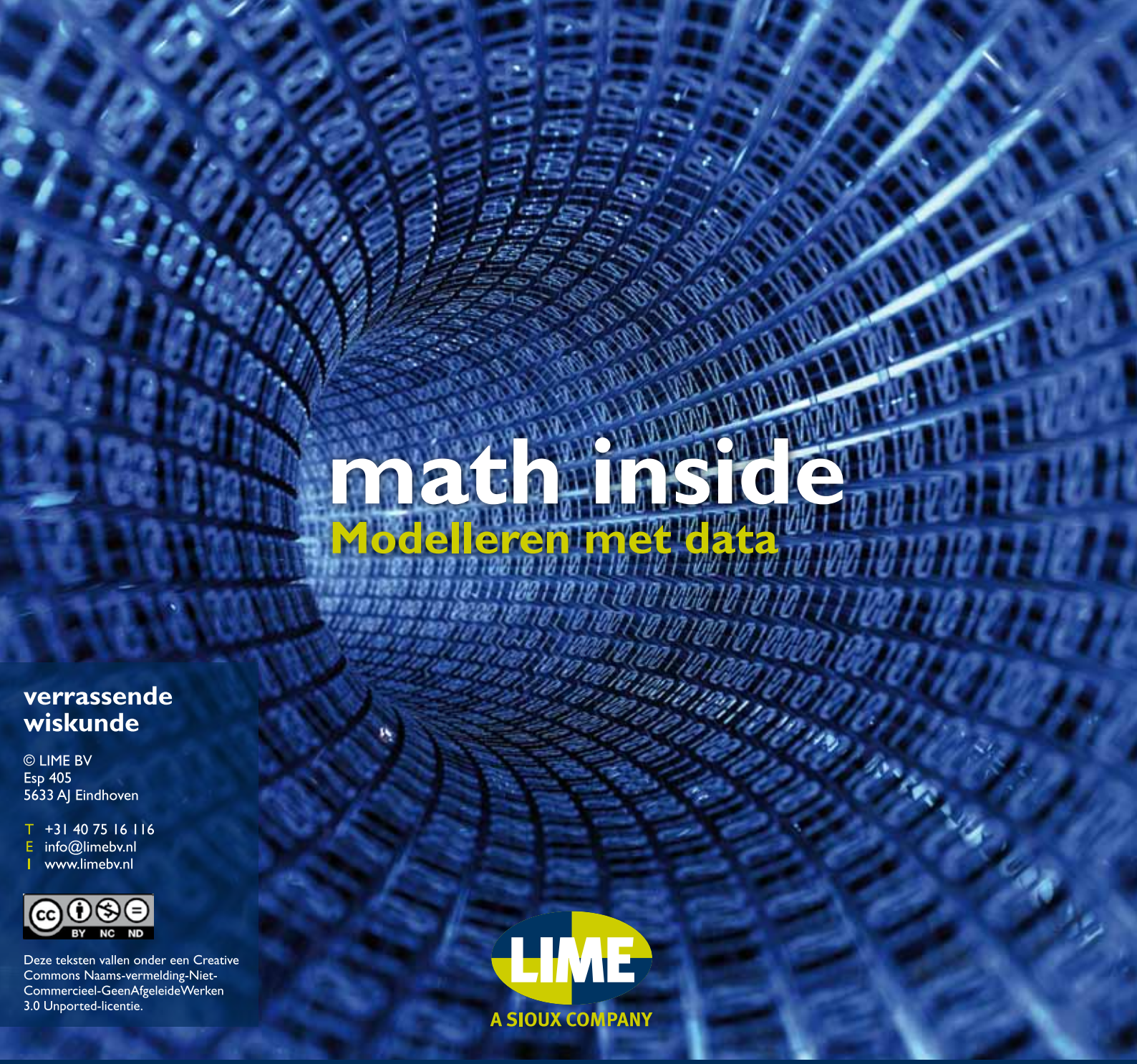


math inside

Modelleren met data

verrassende wiskunde

Esp 405
5633 AJ Eindhoven

T +31 40 75 16 116
E info@limebv.nl
I www.limebv.nl

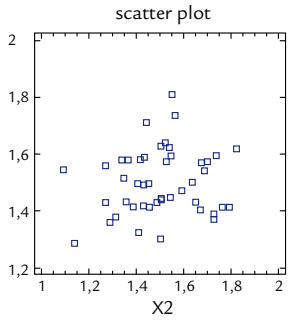
LIME
A SIOUX COMPANY

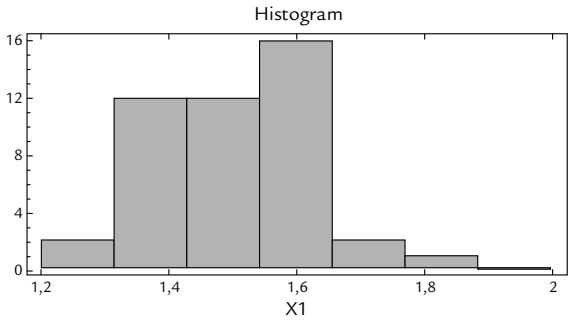
INNOVATION THROUGH COMPUTATION

Modelleren met data

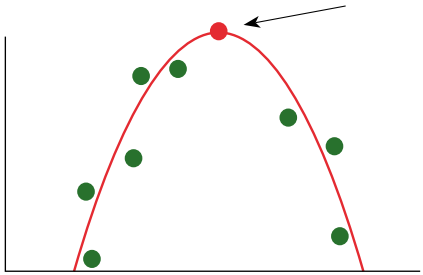
In allerlei bedrijfssituaties worden gegevens bijgehouden, zoals productiegegevens, bijvoorbeeld van de output van productielijnen maar ook van procesparameters en klantgegevens. De stormachtige ontwikkeling van de automatisering heeft er toe geleid dat het in allerlei bedrijfsomgevingen mogelijk is om grote databestanden aan te leggen.

Tegelijkertijd is er een groeiend aantal situaties met weinig data door het steeds meer klantgericht werken met korte productieruns. In beide gevallen kan wiskunde helpen met het halen van nuttige informatie uit deze data. Het meest eenvoudige gebruik van data is het berekenen van eenvoudige statistische kengetallen zoals gemiddelden of standaardafwijkingen. Het is door de overvloed van software gemakkelijk om een eenvoudige grafische weergave van de data te maken zoals een strooidiagram, taartdiagram of histogram. Dit zijn voorbeelden van datareductie, het reduceren van een te hoge complexiteit van data tot meer simpele, laagdimensionale weergave.





Maar dankzij wiskunde kan met de data meer gedaan worden. Voor data-analyse zijn kansmodellen te gebruiken, dat wil zeggen modellen waarvan de uitkomsten van toeval afhangen. Denk bijvoorbeeld aan het werpen van een munt of het aantal klanten in een bepaalde periode. De kansrekening kreeg een degelijk wiskundig fundament door het werk van **Kolmogorov**.



Een simpel voorbeeld om dit te illustreren is het volgende. In het diagram hiernaast zijn metingen weergegeven van een chemisch proces waarin de temperatuur op de horizontale as gevarieerd wordt om tot een maximale opbrengst te komen, aangegeven op de verticale as. Het doel is om te komen tot een zo hoog mogelijke opbrengst. Alleen op de waarnemingen afgaand ligt de maximale opbrengst bij de met rood aangegeven waarneming. Dit wordt vaak het "*pick the-winner*" principe genoemd.

Ervan uitgaande dat de opbrengst een chemisch/fysische verklaring moet hebben, kan geprobeerd worden om op een wiskundig verantwoorde manier een eenvoudige functie te vinden die zo dicht mogelijk bij de waarnemingen blijft. Bij deze methode is er een hoger optimum, aangegeven met de groene pijl. In de praktijk is het probleem natuurlijk veel complexer omdat met veel meer parameters gewerkt wordt.
In de statistiek, het gedeelte van de wiskunde waarin (kans)modellen gemaakt worden aan de hand van data, zijn ogenaamde regressie-analyse modellen een veel gebruikte techniek voor dit type problemen.

Als de data te complex zijn om aan de hand van modellen tot zinvolle uitspraken te komen, dan zijn er wiskundige technieken beschikbaar om tot datareductie te komen. Een voorbeeld hiervan zijn de metingen aan een spectrum, waarin vaak van duizenden golflengten intensiteiten gemeten zijn. Een benaderend model wordt gekozen, waarin alleen de belangrijkste invloedrijke factoren voorkomen. Na het berekenen van een model moet er altijd een validatieslag plaatsvinden.

Een statisticus gebruikt alle data die voorhanden zijn. Data a priori reduceren door voor het gemak gemiddelden te rapporteren is uit den boze. Verder hoeven waarnemingen niet exact bekend te zijn. Zo komt het bij levensduurtesten, waarin door het veranderen van omstandigheden de tijd virtueel versneld wordt, voor dat levensduren van niet gefaalde producten niet bekend zijn. Maar het is wel bekend dat deze producten ten minste de tijd die de test duurde hebben gehaald. Zulke data kunnen en moeten meegenomen worden in statistische analyses. Dit heet gecensureerde data.
Veel methoden, zoals bijvoorbeeld voor data reductie of schattingen, gaan terug op **Fisher**. Bij data mining zijn de data zijn soms zo willekeurig verzameld en vaak zo onnauwkeurig, dat de gebruikelijke statistische modellen niet altijd toepasbaar zijn. Met behulp van voorwaardelijke kansen en de "*regel van **Bayes***" worden dan via de heuristiek verbanden gevonden. Voorbeelden zijn analyses van klantenkaart-gegevens door winkels en kredietrisico-schattingen door banken. De regel van Bayes is een formule om voorwaardelijke kansen te berekenen. Bijvoorbeeld, als men voor verschillend groepen klanten de kans op een financiële transactie kent kan men aan de hand van de data van een klant hiermee de kansen bepalen van die klant, die tot een zekere groep van klanten behoort.