

# Identificação de Comunidades de Dados

## Uma Introdução ao Agrupamento em Risco de Crédito e Marketing

O artigo demonstra como o uso de agrupamento em risco de crédito e *marketing* auxilia na definição de grupos homogêneos de clientes. Por meio dessa distinção, gestores de risco de crédito estão aptos a construir modelos separados para prever inadimplências com maior precisão e gestores de *marketing* podem conceber campanhas publicitárias mais eficazes voltadas diretamente às necessidades específicas de um grupo homogêneo. Apresenta, também, um panorama geral do processo envolvido na determinação de agrupamentos ou nas soluções de segmentação baseadas nas melhores práticas do setor.

Evidentemente, o desenvolvimento de um sistema de agrupamento exige mais do que simples fórmulas estatísticas, equações matemáticas e testes hipotéticos. Para ilustrar alguns conceitos associados à concepção de um sistema de

Susan Alvarez

John Gatschet

Jeffrey Morrison

### Finding Data Communities

#### An Introduction to Clustering in Credit Risk and Marketing

*more accurately predict delinquencies, while marketing managers can design more effective advertising campaigns to speak to the customized needs associated with a homogeneous group. We also provide a general overview of the process involved in determining clustering or segmentation solutions based on industry best practices.*

*It will become apparent that developing a clustering system requires more than statistical formulas, mathematical equations, and hypotheses tests. To illustrate some concepts associated with designing*

*In this article, we introduce how the use of clustering in credit risk and marketing can help companies distinguish homogeneous groups of customers. By making those distinctions, credit risk managers can build separate models to*

*a clustering system while maintaining a level of practicality, we highlight the clustering techniques found in the SAS software system.*

*The design and development of a clustering system is heavily dependent on the nature of the business objective being addressed. For example, analysts handling credit modeling applications often struggle with how to best segment the data in the most homogenous fashion. If there are truly different underlying populations for accounts associated with a new applicant risk model, then building separate models on each cluster or segment should result in a more accurate approach. For marketers, sending the right type of offer, e.g. coupons, discounts, customized advertising, to the right type of customer or market segment should result in higher response rates or reduced customer churn. To set the framework for our recommendations, we've included a step-by-step process that applies to both marketing and credit risk scenario:*

- 1) Preliminary Data Analysis
- 2) Data Transformations
- 3) Variable (Attribute) Selection
- 4) Choosing the Clustering Method
- 5) Determining Optimal Number of Clusters
- 6) Cluster Validation
- 7) Cluster Profiling

agrupamento e, ao mesmo tempo, manter um certo nível de praticidade, foram destacadas as técnicas de agrupamento encontradas no sistema de *software* SAS.

A concepção e o desenvolvimento de um sistema de agrupamento dependem muito da natureza do negócio em questão. Por exemplo, analistas que lidam com aplicações de modelagem de crédito muitas vezes ficam indecisos quanto à melhor maneira de segmentar os dados de forma homogênea. Se houver, realmente, populações subjacentes diferentes para as contas associadas a um novo modelo de risco proposto, construir modelos separados para cada agrupamento ou segmento deve resultar numa abordagem mais precisa. Para a área de *marketing*, enviar a oferta adequada — como cupons, descontos, publicidade personalizada, etc. — a um determinado tipo de cliente ou segmento de mercado, resultará em: maiores taxas de resposta ou menor giro de clientela. Para formar o arcabouço inicial das recomendações, foi incluído um processo passo a passo que se aplica tanto ao cenário de *marketing* quanto ao de risco:

- 1) Análise Preliminar de Dados
- 2) Transformações de Dados
- 3) Seleção de Variáveis (Atributos)
- 4) Seleção do Método de Agrupamento
- 5) Determinação do Número Ideal de Agrupamentos
- 6) Validação de Agrupamentos
- 7) Perfil dos Agrupamentos

## *Análise Preliminar de Dados*

Nunca se começa uma análise de agrupamento rodando um algoritmo “predileto” sobre um conjunto de atributos de dados. O resultado será, inevitavelmente, um conjunto de agrupamentos sem qualquer significado porque praticamente todos os algoritmos de agrupamento são extremamente sensíveis à irregularidade e à escala dos dados. Em vez disso, é preciso desenvolver estatísticas preliminares que descrevam os dados em termos de distribuições, correlações e outras medidas estatísticas simples, como médias, desvios padrão, porcentagem das observações com valores faltantes distribuições de *outliers* e correlações simples entre atributos.

## *Transformações de Dados*

Deve-se iniciar pela eliminação de qualquer atributo fortemente ausente (>30%). Em relação às variáveis remanescentes às quais faltem dados, usa-se um valor *proxy* como o valor médio ou mediano dos dados presentes, ou uma regra que simplesmente estabeleça que o dado vale zero quando apropriado. Em seguida elimina-se qualquer observação extremamente elevada (isto é, os *outliers*), ou fixa-se um teto para as observações na altura do 99º percentil. Finalmente, como os métodos de agrupamento podem ser muito sensíveis a unidades de medida diferentes entre atributos, recomenda-se padro-

## *Preliminary Data Analysis*

Never begin cluster analysis by running a “favorite” clustering algorithm on a set of data attributes. Inevitably, the result will be a meaningless set of clusters because most any clustering algorithm is extremely sensitive to data irregularities and scaling. Rather, preliminary statistics should be developed which describe the data in terms of distributions, correlations, and other simple statistical measures, e.g. determining sample means, standard deviations, percent of observations that have missing values, outlier distributions, and simple correlations among attributes.

O desenvolvimento de um sistema de agrupamento depende da natureza do negócio.

The development of a clustering system is dependent on the nature of the business.

## *Data Transformations*

Begin by deleting any attribute that is highly missing (>30%). For those remaining variables with missing data, use a proxy value like the mean or median value of the non-missing data or use a rule that simply sets the data equal to zero where appropriate. Next, delete any observation that is extremely high, i.e. an outlier, or cap the observations at the 99th percentile. Finally, because clustering methods can be very sensitive to different measurement units across attributes, we recommend standardizing the

data before executing a clustering procedure. A commonly accepted standardization is the z-score:  $(X - X_{Mean}) / X_{SDev}$ . Another option is:  $X_{Std} = X / \sqrt{\text{Maximum}(X) - \text{Minimum}(X)}$ . SAS automatically performs either of these data transformations with a procedure called PROC STANDARD. Standardizing your data also can help identify, remove or cap extreme values. A general rule of thumb is that the values of the standardized variables should not exceed +/-5 standard deviations. Therefore in practice, we suggest analysts perform a quick standardization process on the data as part of the preparation phase.

### Variable (Attribute) Selection

Regardless of whether the business need is marketing or credit risk related, variable selection is one of the most important phases of cluster development. For regression models, variable selection is not as crucial. Regression software now has built in features such as "stepwise selection" that weed out unimportant variables by using tests that are designed around long-standing statistical foundations. This is not the case for cluster solutions because there is no target or dependent variable to use and the assumptions related to most traditional statistical tests are violated. Moreover, studies show that including even a single irrelevant variable in a cluster solution can have a dramatic impact on the end result. Given today's analyst is typically faced with hundreds of variables to choose from, finding the 10 or 20 best sui-

nizar os dados antes de realizar um procedimento de agrupamento. Uma padronização comumente aceita é o z-score:  $(X - X_{Média}) / X_{Dpad}$ . Outra opção é:  $X_{pad} = X / \sqrt{\text{Máximo}(X) - \text{Mínimo}(X)}$ . O SAS faz automaticamente qualquer uma dessas transformações por meio de um procedimento chamado PROC STANDARD. Padronizar os dados também ajuda a identificar, eliminar ou fixar teto para os valores extremos. Uma boa regra geral é a de que os valores das variáveis padronizadas não devem superar +/-5 desvios padrão. Assim, na prática, sugere-se aos analistas que realizem um rápido processo de padronização dos dados como parte da fase preparatória.

### Seleção de Variáveis (Atributos)

Independentemente da necessidade do negócio estar relacionada a *marketing* ou risco de crédito, a seleção de variáveis é uma das fases mais importantes do desenvolvimento de agrupamentos. Para modelos de regressão, a seleção de variáveis não é tão crucial. Os softwares de regressão hoje dispõem de características embutidas como a "seleção *stepwise*", que eliminam variáveis sem importância por meio de testes concebidos em torno de bases estatísticas tradicionais. Não é o caso das soluções de agrupamentos porque não há variável-meta ou dependente que possa ser usada e as premissas relacionadas aos testes estatísticos mais tradicionais são violadas. Ademais, estudos demonstram que a inclusão até mesmo de uma só variável irrelevante numa solução de agrupamento pode ter efeito dramático sobre o resultado final. Dado que o analista de hoje se depara habitualmente com centenas de variáveis, identificar as 10 ou

20 mais adequadas para desenvolver um sistema de agrupamento pode ser um desafio.

Uma solução popular para enfrentar o desafio da seleção de variáveis é a análise fatorial ou, mais especificamente, a análise de principais componentes (APC). A análise fatorial reduz a duplicação de variáveis, criando um conjunto de combinações lineares que permite ao analista decompor as informações em suas dimensões ou seus fatores fundamentais.

Embora a verdadeira análise fatorial admita a existência de elementos fundamentais num conjunto de dados ou por trás dele, a análise de principais componentes é mais especializada. Procura apenas explicar a variabilidade do conjunto. Em suma, a APC é uma técnica de redução de dados que permite ao usuário resumir o conteúdo informacional expresso num grande número de variáveis com um número substancialmente menor de componentes principais sem, com isso, perder informações críticas.

O resultado da APC apresentará diversos componentes principais ou “fatores”. Eis como funciona o processo, em linhas gerais: deriva-se um conjunto de transformações lineares ou ponderações de forma que o primeiro fator explique a maior parte da variabilidade dos dados. Isso está representado na Figura 1, um exemplo simples em que há apenas duas variáveis. A linha azul representa o eixo associado ao primeiro componente principal, que contém a maior parte da variação entre as duas variáveis. O segundo fator, representado pelo eixo da linha vermelha, explica a maior agrupamento de variáveis como resultado do que explicou o primeiro fator. Observe-se que

*ted for developing a clustering system can be challenging.*

*One popular solution to address the variable selection challenge is factor analysis, or more specifically principal component analysis (PCA). Factor analysis reduces variable duplication by creating a set of linear combinations that allows the analyst to collapse the information into underlying dimensions or factors.*

*While true factor analysis assumes underlying “factors” exist in or “behind” a dataset, principal component analysis is more specialized. It merely attempts to account for the variability in the dataset. In short, PCA is a data reduction technique that allows a user to summarize the information content expressed by a large number of variables with a substantially smaller number of principal components without losing critical information.*

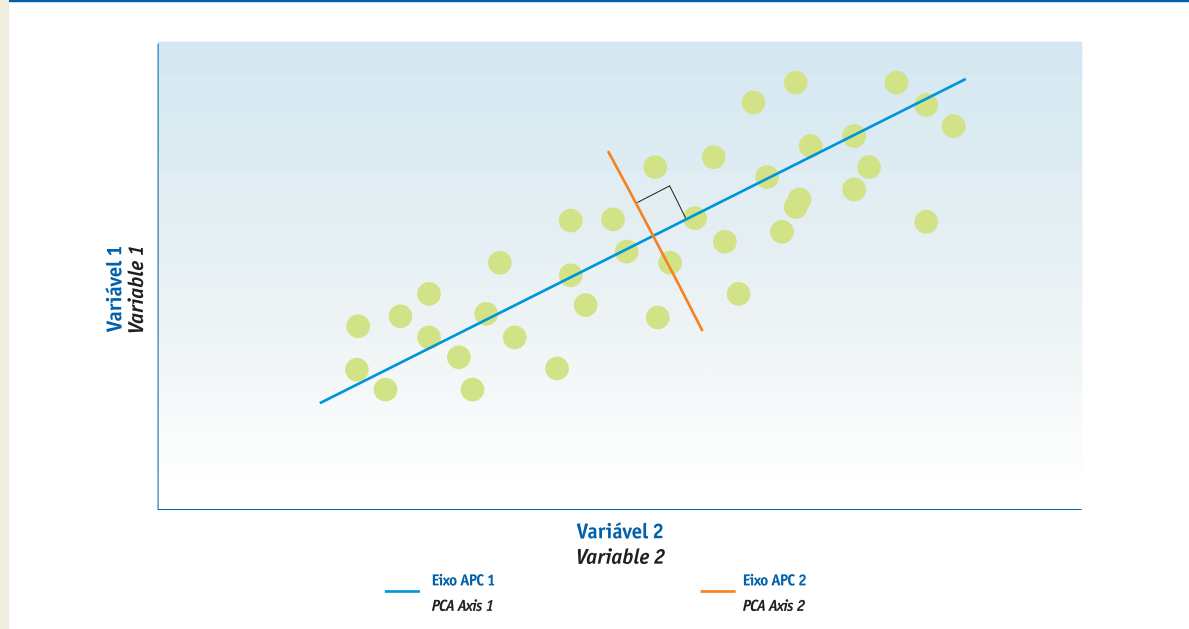
*The output of PCA is a number of principal components or “factors”. At a high level, here’s how the process works. A set of linear transformations or weights is derived such that the first factor explains the greatest amount of variability in the data. This can be seen in Figure 1 below, a simple illustration where there are only two variables. The blue line represents the axis associated with the first principal component, which contains the majority of the variation between the two variables. The second factor, represented by the red line axis, explains the next greatest amount of variability given what was explained by the first factor. Note that*

*the geometric angle between these axes is exactly 90 degrees, which means they are not correlated with one another.*

o ângulo geométrico entre os dois eixos é de exatamente 90°, o que significa que não há correlação entre eles.

Figura 1

Figure 1



*Based on this iterative process, if you have 100 original variables, you would end up with 100 principal components with the later components explaining less and less of the data variation. Often it takes only a small number of principal components, i.e. five or six, to explain 95 percent of the variation in the data, even if you have 100 or more original variables.*

*Factor analysis as well as principal components also produce something called factor loadings. A factor loading reflects the correlation between the principal component and the original variables. These loadings range in value from -1 to 1. Therefore, if you have high correlation numbers in the first factor loading for the original variables, e.g. account age, balance, and credit utilization, then those variables comprise the majority of the variation*

Com base nesse processo iterado, se tivermos 100 variáveis originais, acabaremos com 100 componentes principais, sendo que os últimos desses explicam o mínimo da variação dos dados. Muitas vezes basta um pequeno número — cinco ou seis — de componentes principais para explicar 95% da variação dos dados, mesmo que haja 100 ou mais variáveis originais.

A análise fatorial, assim como a de componentes principais, produz algo chamado cargas fatoriais. Uma carga fatorial reflete a correlação entre o componente principal e as variáveis originais. O valor dessas cargas vai de -1 a 1. Assim, se houver elevada correlação na primeira carga fatorial das variáveis originais, como idade da conta, saldo e utilização de crédito, essas variáveis representarão a maior parte da variação do primeiro componente principal. Quanto mais

próxima de 1 for a carga fatorial de uma variável específica qualquer, maior será a relevância da variável para a representação do fator. Considerações como correlação com outras variáveis e a predominância de valores ausentes também devem ser ponderadas antes de se escolher qualquer variável para representar um fator.

### *Escolha do Método de Agrupamento Adequado*

Após reduzir o subconjunto original de variáveis para que atinja um tamanho administrável (de 10 a 20), o passo seguinte será determinar o algoritmo de agrupamento correto. Tradicionalmente, a análise de agrupamentos se divide em dois tipos: a hierárquica e a não-hierárquica, também chamada de média k. Há dois subtipos de análise hierárquica: a aglomerativa e a divisiva. Uma análise de agrupamentos hierárquica aglomerativa começa com cada observação formando seu próprio agrupamento. Em seguida, combinam-se os agrupamentos que estejam mais próximos. O processo continua até que todas as observações formem um só agrupamento. Uma análise de agrupamentos hierárquica divisiva parte de um só agrupamento e avança até chegar ao número de observações. Para o agrupamento aglomerativo o SAS oferece diversos métodos para definir como combinar os agrupamentos mais próximos.

No SAS, muitas vezes prefere-se o PROC FASTCLUS ao PROC AGRUPAMENTO porque o primeiro exige menos tempo. Mas uma das limitações do procedimento consiste na criação de apenas uma só solução específica, isto é, com um número pré-determinado de agrupamentos. Na prática, recomendamos que os analistas sub-

*for that first principal component. The closer a factor loading for a specific variable is to 1, the greater relevance that variable has in representing that factor. Considerations such as correlation to other variables and the prevalence of missing values also must be weighed prior to choosing any variable to represent a factor.*

### *Selecting the Appropriate Clustering Method*

*After the subset of original variables is reduced to a manageable set (10-20), the next step is determining the correct clustering algorithm. Traditionally, cluster analysis is broken into two types: hierarchical and nonhierarchical, which is also called k-means. Within hierarchical, there are two subtypes: agglomerative and divisive. An agglomerative hierarchical cluster analysis starts with each observation forming its own cluster. It then combines clusters, which are closest in distance. This process continues until all observations form a single cluster. A divisive hierarchical cluster analysis starts with one cluster and works back to the number of observations. For agglomerative clustering, SAS offers a number of different methods of defining how closest clusters are combined.*

*In SAS, PROC FASTCLUS is often preferred over PROC CLUSTER because it requires less computing time. However, one limitation of the procedure is that it only creates a single specific solution, i.e. one with a preset number of clusters. In practice, we*

*recommend analysts submit a variety of cluster specifications in a trial and error fashion before deciding on the final clustering structure.*

### **Determining Optimal Number of Clusters**

*As stated, if you are using PROC FASTCLUS in SAS, then you must program how many clusters it has to create. Although this seems somewhat judgmental, there are some general rules of thumb that could help along with a few statistical measures. For example, if you are new to the market and want to capture a general representation of the risk or marketing dynamics, you may consider designating four to eight clusters. On the credit side, a clustering system consisting of more than 15 homogeneous groups means that you have to build quite a few risk models or scorecards to implement and support. SAS has three statistical measures reflecting "goodness of fit" that can guide you on how to better refine this number: Pseudo F-statistic, Cubic Clustering Criterion, and R-squared. In general, the higher these measures are, the better your clustering solution. In reality, almost all clustering systems are developed by selecting a variety of starting values for the number of clusters. This is analogous to a credit risk or marketing analyst building a spectrum of regression models before selecting the best one.*

*There are other ways to pick the optimal number of clusters. One of the most re-*

*metam diversas especificações de agrupamentos por tentativa e erro antes de optar pela estrutura de agrupamento definitiva.*

### **Determinação do Número Ideal de Agrupamentos**

Como foi visto, quando se usa o PROC FASTCLUS do SAS, é preciso indicar quantos agrupamentos devem ser criados. Embora isso possa parecer um tanto subjetivo, há algumas regras gerais que podem ajudar com algumas medidas estatísticas. Por exemplo, quando uma empresa for nova no mercado e quiser capturar uma representação geral da dinâmica do risco ou do mercado, pode considerar de quatro a oito agrupamentos. Do lado do crédito, um sistema de agrupamento que consista em mais do que 15 grupos homogêneos significa que será necessário construir diversos modelos de risco ou *scorecards* que precisarão ser implementados e suportados. O SAS traz três medidas estatísticas que refletem a adequação do encaixe ("*goodness of fit*") e podem ser usadas para melhor refinar esse valor: Pseudo Estatística F, Critério de Agrupamento Cúbico e R-Quadrado. De modo geral, quanto maiores essas medidas, melhor a solução de agrupamento. Na realidade, quase todos os sistemas de agrupamento são desenvolvidos por meio da seleção de diversos valores iniciais para o número de agrupamentos. Isso é análogo a um analista de risco de crédito ou de *marketing* que constrói diversos modelos de regressão antes de escolher o melhor deles.

Há outras maneiras de escolher o número ideal de agrupamentos. Um dos mais recentes avan-

ços tecnológicos refere-se ao apoio dado pelo emprego do PROC MODCLUS do SAS3. Ao especificar um conjunto de parâmetros de suavização (“smoothing parameters”), ou seja, 5, 8, 10, 12, 15, o procedimento classificará algumas observações e não outras, fornecendo ao mesmo tempo, automaticamente, o número de agrupamentos de cada conjunto de soluções. Basta eliminar as soluções de agrupamento que não classifiquem bem todos os dados.

Há mais o que considerar ao definir a especificação final de agrupamento. Por exemplo, o tamanho de cada agrupamento é importante. De modo geral, na maioria das soluções de agrupamento há um agrupamento grande e diversos menores. Se o maior agrupamento representar parcela grande demais da população, o objetivo do agrupamento estará sendo ignorado. Se o número de componentes de um agrupamento for pequeno demais, embora ele possa identificar bem um segmento específico, o agrupamento poderá ser pequeno demais em termos práticos. Outra consideração importante: há em cada agrupamento uma amostra adequada para a realização de modelagem posterior ou a execução de estratégias de implementação?

A Figura 2 mostra como o procedimento CANDISC do SAS pode apresentar a solução de agrupamento sob forma gráfica. Quando há milhares de observações, pode ser desejável tomar uma amostra aleatória menor para fins de visualização. O procedimento CANDISC, semelhante ao de principais componentes, é um meio simples de ver como a solução separa os dados em diferentes agrupamentos por meio da criação de dois eixos que representam o poder de separação dos

*cent technological developments that can help is by using SAS' PROC MODCLUS procedure<sup>3</sup>. By specifying a set of smoothing parameters, i.e. 5, 8, 10, 12, 15, the procedure will classify some observations and not classify others, while automatically providing you with the number of clusters in each solution set. Simply eliminate those cluster solutions that do not do a good job in classifying all your data.*

*There are other considerations to keep in mind when deciding the final cluster specification. For example, the size of each cluster is important. Generally, in most clustering solutions, there is one large cluster and several smaller clusters. If the largest cluster represents too much of the population, then the clustering's objective is ignored. If a cluster membership is too few, while it might identify a specific segment well, practically speaking, the segment may be too small. Another important consideration: Is there an adequate sample within each cluster to perform any follow-up modeling or execute implementation strategies?*

*Figure 2 shows how the CANDISC procedure in SAS can display the cluster solution in graphical form. If you have thousands of observations, you may just want to take a smaller random sample for display purposes. The CANDISC procedure, similar to principal components provides a simple method to see how your cluster solution separates your data into different clusters by creating two axes, which represent the*

separation power of your data.

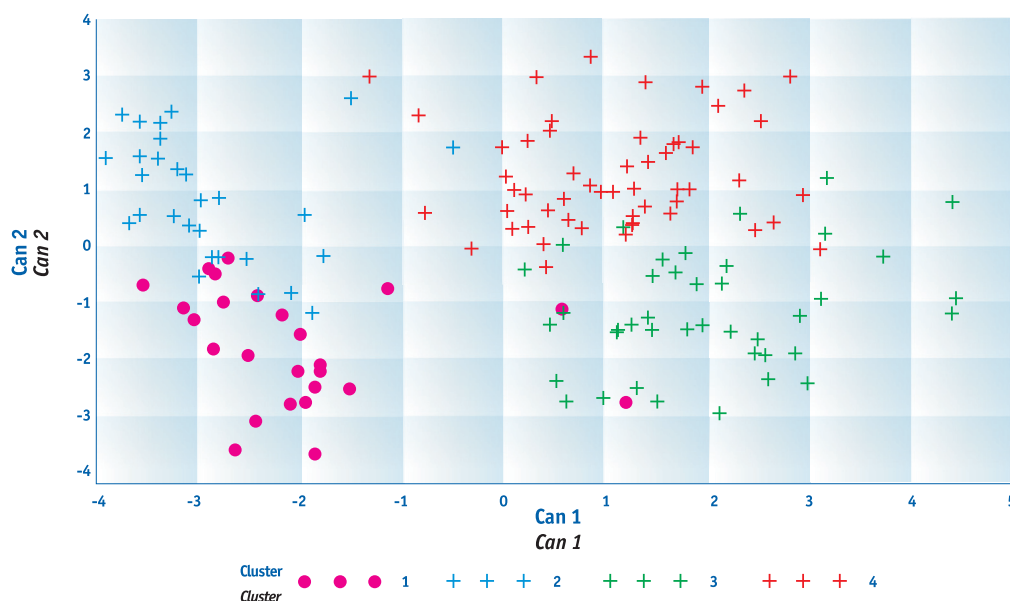
The more separation you get between clusters, the better your cluster solution. Finally, it is not uncommon to make some manual modifications to the final clustering solution based on the additional insight revealed during the profiling. One such modification might be manually combining two small clusters originally split on one attribute. Another is to manually force a split of a larger cluster on a predefined attribute.

dados.

Quanto maior a separação entre agrupamentos, melhor será a solução. Finalmente, não é raro fazer algumas modificações manuais da solução de agrupamento final com base nos *insights* adicionais revelados durante a visualização do perfil. Uma modificação poderia ser a combinação de dois pequenos agrupamentos originalmente separados por conta de um atributo. Outra é forçar manualmente a divisão de um agrupamento maior com base em algum atributo predeterminado.

Figura 2

Figure 2



## Validation

Unlike credit scoring and marketing response models, there are no hard and fast rules about validating your cluster solution. One approach is to randomly split your data into a development sample and a holdout sample, just as you would in a regression approach. By performing

## Validação

Ao contrário dos modelos de *credit scoring* e de resposta de *marketing*, não há regras claras e inflexíveis sobre a validação da solução de agrupamento. Uma abordagem é a divisão aleatória dos dados entre uma amostra de desenvolvimento e outra de reserva, como se faria com uma abordagem de regressão. Realizando as mesmas

rotinas de análise de agrupamentos nas duas amostras independentemente, deveremos obter resultados semelhantes se a solução estiver corretamente validada.

### *Perfil dos Agrupamentos*

Embora útil nas aplicações de risco de crédito, é fundamental para as estratégias de segmentação de *marketing* que seja atribuída a cada grupo homogêneo do sistema de agrupamento uma descrição adequada e sensata. Isso ajuda a fazer com que os segmentos de mercado “ganhem vida”, com rótulos como urbanos sofisticados ou informados sobre crédito. Isso normalmente se faz olhando para as estatísticas básicas, isto é, médias, frequências e porcentagens, de diversas variáveis de seus dados, algumas das quais podem não ter estado envolvidas na solução de agrupamento final. Uma maneira mais formal de identificar candidatas ao perfil poderia ser usar uma abordagem por classificação como a análise múltipla discriminada, PROC DISCRIM do SAS, para fazer emergir as variáveis que mais ajudam a separar os agrupamentos. Quanto maiores as diferenças entre esses perfis de variáveis, mais fácil será atribuir rótulos sensatos que descrevam as características dos diversos segmentos.

### *Últimas Observações*

Desenvolver uma solução de agrupamento para segmentação de mercado e risco de crédito é mais arte do que ciência. Uma das principais diferenças entre a regressão e a análise de agrupamento é a ausência de testes estatísticos para ajudar a escolher a “melhor” solução. Na regressão, se não houver conteúdo informacional nos

*the same cluster analysis routines on both samples independently, you should get similar results provided your solution validates properly.*

### *Cluster Profiling*

*Although useful in credit risk applications, it is imperative in a marketer's segmentation strategies that adequate and sensible descriptions are assigned for each homogenous group in the clustering system. These assignments help make the market segments “come alive” with labels like urban sophisticates or credit savvy. Typically this is accomplished by looking at basic statistics, i.e. means, frequencies, percentages, for a variety of variables in your data, some of which were not involved in the final cluster solution. A more formal way to identifying profile candidates might be to use a classification approach such as multiple discriminate analysis, SAS'PROC DISCRIM, to bring to the surface those variables best helping to separate the clusters. The greater the differences in these variable profiles, the easier sensible labels can be assigned describing the characteristics of the various segments.*

### *Concluding Remarks*

*Developing a clustering solution for market segmentation and credit risk is more of an art than a science. One of the main differences between regression and cluster analysis is the lack of statistical tests to help select the “best” solution. In regression, if there is no information content in your data, then you*

*cannot build a model with statistically valid predictors. However, in cluster analysis, the nature of the procedures will always guarantee some cluster solution. If you want five clusters, you can receive five clusters. This typically makes the development of a cluster scheme for credit risk or marketing more ad hoc in nature. In fact, twenty analysts, depending on their background and experience, could develop twenty different clustering schemes with the same data and none of them would necessarily be wrong. Testing and comparing clustering solutions involving alternate sets of clustering variables and varying number of clusters can give the analyst some measure of the relative robustness of the preferred solution.*

*Given the flexibility that cluster analysis offers, it is important to take a structured approach to meet your business objective. Often the preliminary analysis step is overlooked and the proper scaling and outlier detection is not given the attention it deserves. It is important that care be taken in the variable selection step as the introduction of just a single erroneous variable can drastically alter the cluster solution. However, at the end of the day, the most important aspect of the*

*dados, não será possível construir um modelo de preditivas estatisticamente válidas. Na análise de agrupamento, contudo, a natureza do procedimento garante sempre alguma solução. Se quisermos cinco agrupamentos, podemos receber essa quantidade. Isso normalmente faz com que o desenvolvimento de um esquema de agrupamento para fins de risco de crédito ou *marketing* seja mais subjetivo. De fato, vinte analistas, dependendo de seu histórico e sua experiência, poderiam desenvolver, a partir dos mesmos dados, vinte esquemas de agrupamento diferentes sem que qualquer um deles estivesse necessariamente errado. Testar e comparar soluções de agrupamento que envolvam conjuntos alternados de variáveis de agrupamento e um número variado de agrupamentos pode conferir ao analista alguma medida da robustez relativa da solução preferida.*

*Tendo em vista a flexibilidade que a análise de agrupamento oferece, é importante adotar uma abordagem estruturada para que se atinja o objetivo empresarial. Muitas vezes, a etapa de análise preliminar é desprezada e, com isso, não se dá a devida atenção à escala correta e à detecção de *outliers*. É importante ter cuidado na etapa de seleção de variáveis, uma vez que a introdução de uma só variável errônea pode alterar drasticamente a solução de agrupamento. Mas, no fim*

*É preciso desenvolver estatísticas preliminares que descrevam os dados.*

*Preliminary statistics should be developed which describe the data.*

das contas, o aspecto mais importante do esforço de agrupamento é entender como o esquema de segmentação será usado em relação ao objetivo empresarial. Do lado do risco, um excesso de agrupamentos irá resultar na construção de um número excessivo de *scorecards* a serem suportados e mantidos. Se os segmentos não distinguirem efetivamente entre grupos homogêneos relevantes, ter um *scorecard* para cada agrupamento aumentará a precisão global da minimização do risco. Com referência a *marketing*, se os agrupamentos não puderem ser traduzidos em segmentos significativos que possam ser implementados em campanhas promocionais ou publicitárias, sua utilidade será limitada.

### Referências

- 1). HAIR, Joseph, e ANDERSON, Rolphe. *Multivariate Data Analysis*, 4a edição. Copyright 1995, Prentice-Hall, Inc.
- 2). VRIENS, Marco. *Market Segmentation, Analytical Developments and Application Guidelines*. Março de 2001. Technical Overview Series.
- 3). THOMPSON, Mark E. *The Science and Art of Market Segmentation Using PROC FASTCLUS*. Paper 270. Forefront Economics, Inc.

© Jeffrey S. Morrison é Gerente Sênior da TransUnion, LLC em Atlanta, Geórgia, onde lidera a função de Pesquisa e Desenvolvimento de análises e trata de projetos especiais de Econometria. Jeffrey formou-se pelo Georgia Institute of Technology, com graduação em Economia e Administração, e obteve seu mestrado em Economia Empresarial pela Georgia State University. Sua carreira profissional de análise abrange diversos setores, inclusive distribuição de gás natural, telecomunicações, crédito ao consumidor e bancos comerciais e de varejo. Jeffrey proferiu palestras em diversas conferências sobre modelagem estatística e previsão em todos os Estados Unidos e já publicou mais de 30 artigos em periódicos especializados. Entre em contato com Jeff no endereço [jmorrison@transunion.com](mailto:jmorrison@transunion.com)  
John Gatschet é Consultor Técnico Sênior da TransUnion

*clustering effort is understanding how the segmentation scheme is going to be used in relationship to the business objective. On the risk side, too many clusters will result in building too many scorecards that have to be supported and maintained. If the segments really don't distinguish between relevant homogeneous groups, then having additional scorecards for each cluster will not increase overall accuracy in mitigating risk. On the marketing side, if the clusters cannot be translated into meaningful segments that can be implemented in promotional or advertising campaigns, then their usefulness is limited.*

### References

- 1). HAIR, Joseph, and Anderson, Rolphe. *Multivariate Data Analysis*, 4th edition. Copyright 1995, Prentice-Hall, Inc.
- 2). VRIENS, Marco. *Market Segmentation, Analytical Developments and Application Guidelines*. March 2001. Technical Overview Series.
- 3). THOMPSON, Mark E. *The Science and Art of Market Segmentation Using PROC FASTCLUS*. Paper 270. Forefront Economics, Inc.

© 2005 RMA. Jeffrey Morrison is Senior Manager of Analytics for TransUnion where his role includes leading the Research and Development initiatives of the organization as well as special projects in Econometrics. Jeffrey graduated from Georgia Institute of Technology with degrees in Economics and Management and then earned a Masters of Science in Business Economics from Georgia State University. His professional career in analytics spans a number of industries including Natural Gas Distribution, Telecommunications, Consumer Credit, and Commercial & Retail Banking. Jeffrey has spoken in numerous conferences throughout the United States on subjects related to statistical modeling and forecasting, and has published over 30 articles in applied Journals. Jeff can be contacted at [jmorrison@transunion.com](mailto:jmorrison@transunion.com)

*John Gatschet is Senior Technical Consultant for TransUnion and has nearly 20 years of experience in the design, development and implementation of marketing and credit risk models. John also ran a consulting business that focused on monitoring and managing credit scoring models for banks and large credit unions. Beyond his experience in traditional models, John has also developed transaction-based models using regression techniques, neural networks, and decision trees to predict fraud and consumer nonpayment. John holds a M.S. degree in Statistics from Kansas State University, and a B.A. degree in Mathematics (summa cum laude) from Saint Louis University with minors in Computer Science and German. John can be contacted at [jgatsch@transunion.com](mailto:jgatsch@transunion.com).*

*Susan Alvarez is Senior Manager of Analytical with over 20 years of experience in statistical modeling and data analysis/mining for financial services, insurance, publishing, telecommunications, catalog and continuity clubs. Susan has extensive experience building both customized and generic consumer segmentation systems. She has also developed scoring models to predict consumer bankruptcy, credit and insurance risk, response/conversion, cross-sell, and customer value and retention. Susan holds a M.S. degree in Management/Market Research from the School of Management at the Georgia Institute of Technology and a B.A. degree cum laude in Economics from Mount Holyoke College. Susan is fluent in Spanish and can be contacted at [smalvar@transunion.com](mailto:smalvar@transunion.com)*

*e tem quase 20 anos de experiência em desenho, desenvolvimento e implementação de modelos de risco de marketing e de crédito. John também administrou uma empresa de consultoria que se concentrava no monitoramento e gerenciamento de modelos de credit scoring para bancos e grandes co-operativas de crédito. Além de sua experiência com modelos tradicionais, John também desenvolveu modelos transacionais baseados em técnicas de regressão, redes neurais e árvores de decisão para prever fraudes e não-pagamento por parte de clientes. John é Mestre em Estatística pela Kansas State University, e Bacharel em Matemática (summa cum laude) pela Saint Louis University, com graduação secundária em Ciências da Computação e Alemão. John pode ser contatado no endereço [jgatsch@transunion.com](mailto:jgatsch@transunion.com)*

*Susan Alvarez é Gerente Sênior de Análises e tem mais de 20 anos de experiência em modelagem estatística e análise/mineração de dados para serviços financeiros, seguros, editoras, telecomunicações e clubes de compras e fidelidade. Susan tem larga experiência na construção de sistemas de segmentação de consumidores, sejam customizados ou genéricos. Ela também desenvolveu modelos de scoring para prever insolvência de consumidores, risco de crédito e de seguro, resposta/conversão, venda cruzada e valor e retenção de clientes. Susan é Mestre em Administração/Pesquisa de Mercado pela Escola de Administração do Georgia Institute of Technology e Bacharel (cum laude) em Economia pelo Mount Holyoke College. Susan é fluente em espanhol e pode ser contatada no endereço [smalvar@transunion.com](mailto:smalvar@transunion.com)*