Relatório de Trabalho Final
Computação Estatística II (22009)

Ricardo Neves Pires, nº estudante: 2001645
2001645@estudante.uab.pt

Visualizing statistical models: Removing the blindfold

Hadley Wichkam, Dianne Cook and Heile Hofmann

12 de junho de 2021

Índice

Resumo: O intuito do seguinte documento remete-se à analise e critica do artigo cientifico de Wickham, Cook e Hofmann na abordagem, estratégia e processo dos métodos de visualização de modelos estatisticos . Isto é, os autores dão enfâse não propriamente à visualização dos dados, algo que acontece por norma em primeira instância no momento de exploração de um probelma ou base de dados, contudo não limitante a essa fase mas sim à visualização do que sucede à posteriori. Visualizar o que sucede na fase de modelação, como os modelos podem ser visualizados e as diversas forma de obter uma precisão e adequação dessa mesma visualização sobre os modelos. Passa-se por um procedimento de execução e análisa do modelo em termos visuais que possibilita melhorias na construção, diagnóstico e sumarização do mesmo.

Palavras-chave: model visualization, exploratory data analysis, data mining, classification, high-dimensional data, networks.

1. Introdução

A forma mais simples de caracterizar a razão para o uso de visualizações de modelos remete-se ao facto das pessoas serem muitas vezes sobrecarregadas com dados. A capacidade de transformar os dados em informações que eventualmente poderão ser usadas na tomada de decisões, melhorias de produtos, aumento na compreensão de problemas, etc, é uma tarefa nem sempre fácil e até por vezes enorme. A componente visual apresenta o caminho mais amplo para as nossas capacidades cognitivas e chamada de atenção, seja na detenção fácil de anomálias ou padrões, e é por isso que enfatizar em técnicas gráficas interativas e dinâmicas tem sido ao longo de décadas um ponto fulcral de progresso (Liu, 2017). A complexidade e o crescimento da quantidade de dados, necessita que novas abordagens e *frameworks* sejam implementadas com melhorias constantes, tornado-se um desafio, seja pela limitante computacional que possa existir, quer pelo paradigma existente e enraizado. Felizmente, ao longo dos últimos anos o poder das máquinas e capacidade de processamento tem evoluido e crescido a largos passos. Contudo, os problemas de dados com dimensões elavadas é desde longa data e até aos dias de hoje um problema constante de debate, investigação e de dificil resolução. A escolha de visualizações necessita algum cuidado e considerações a ter em conta à priori. Os autores afirmam que "converting datavis to model-vis is not always straightforward.". Porém, ao longo do artigo a ideia subjacente é muito contrária a esta afirmação e o leitor ficar em parte com a ideia que a criação de visualizações é bastante fácil, acessível e directa. Talvez o seja no contexto em questão, pois é evidente que as bases de dados usadas são bastante amigas e de manipulação fácil. Em situações reais e em muitos conjuntos de dados contemporaneos, o número de observações é relativamente pequeno em comparação com o número de variáveis (high dimension low sample size) o que leva a muitos problemas, nomeadamente com uso de métodos classicos. Wickham (2011), elabora de forma astuta e no eco-sistema R as chamadas *tour* para *high-dimensional data* , incorporando diversas projeções no espaço. Porém, não soluciona tudo com esta abordagem. Alguns dos conceitos discutidos no artigo são bastante comuns na análise de dados (e.g. visualizar as fronteiras na regra de decisão) e outros como grand tours que por norma são pouco usados. Frequentemente a intuição entre praticantes sobre a visualização de dados é comum e baseada em experiências ad hoc. O artigo sintetiza de forma coerente novas abordagens possíveis, fornece exemplos de novas visualizações que expõem recursos especificos de modelos complexos que poderão eventualmente ser metido ao uso dos praticantes/analistas. A permissa de Wickham e colegas é a de que, exibir mais dados é melhor do que exibir menos. Pese embora que ao longo do artigo somos levados à questão, "Qual o objetivo de visualizar modelos?", sem propriamente obter

uma resposta evidente.

São referidas três estrategias ou princípios basilares no artigo, uma é a de visualizar o modelo no espaço dos dados (model in data space, m-in-ds) e os dados no espaço do modelo (data in model space, d-in-ms). Afirmam que visualizar o modelo no espaço dos dados em dimensões superiores permite obter uma sumarização dos dados obtido pelo modelo e em seguida entender melhor o ajuste do modelo. Outra estatégia é a de explorar multíplos modelos e não apenas um só. Com isto pretende-se obter mais informação sobre os dados tal como sucede com várias estatisticas sumárias são mais informativas do que somente uma. Por último, explorar os diversos modelos, visualizar o processo de ajustamento do modelo e olhar para a iterações produzidas por forma a entender como os dados contribuiem para o modelo final. Ao longo das diversas secções são exemplificados e incorporadas estas estratégias em questão. A secção 3 tem como exemplo a MANOVA, modelos de classificação e agrupamento hierárquico para ilustrar o primeiro princípio, confrontando m-in-ds com os d-in-ms. A estratégia de explorar multiplos modelos é realçada na seção 4, com o exemplo de modelos lineares. Dá-se enfâse em visualizar uma coleção de vários modelos e como pode ser vantajoso. A parte sobre visualizar o processo de ajustamento do modelo é delineada na secção 5 com o exemplo de *self-organizing maps.* Por fim os autores tentam na secção 6 aplicar as três estratégias num só exemplo e visualizar de forma interativa uma rede neuronal.

2. Os visuais e modelos

O cerne deste artigo prende-se com a visualização de uma coleção de modelos e o ajuste de modelos de forma iterativa. Isto é, maximar o número de gráficos dos dados com sobreposição dos modelos ajustados. Fica-se pouco escalerecido de que forma e com que ferramentas possibilitam a própria organização dessa mesma coleção de modelos por forma a facilitar a comparação e análise de cada modelo. Visualizar uma coleção de modelos não leva forçosamente a uma melhoria de modelos ajustados ou inferência, nem talvez a forma mais eficaz e eficientes de o fazer. Certo é, que em muitos casos o uso de gráficos simples, usuais e de leitura fácil consigam trazem o mesmo entendimento e resultados que as abordagens sugeridas pelos autores. Sem dúvida que os autores tentam deixar um guia ou sugestão de boas práticas segundo o que entendem. Há portanto uma estrutura ou procedimento possível de seguir e dessa forma permitir a outros trabalhar fora do ad hoc, com iniciativa de criação de conhecimentos mais aprofundados e detalhados sobre os problemas em mão, e progredir com melhores visualizações e técnicas visuais. O objetivo típico no dia-a-dia de um analista é o uso

de model vis para verificação e ajustar modelos antes da sumarização. Não fica propriamente evidente que as propostas dos autores sejam adequadas para a correção de modelos. Ficando em parte a questão em aberto. A forma que a adopção intuittiva e implementação no dia a dia de model vis deste gênero sobre questões complexas seja tão fácilmente usada quanto é percepcionado no artigo pode ser enganadora. Posto isto, os autores fornecem um background do que irão ser os métodos da sua tese. Os autores entram no detalhe sobre o termo e distinção de modelo. Isto é, referem que existem três niveis de especificidade, a familia do modelo, a forma do modelo e o modelo ajustado. Em parte a justifcação para tal vem referido do seguinte modo, *"Knowing the model family and form does not guarantee we know what the fitted model looks like. For example, the parameters of the linear model can be interpreted as the change in the response when the predictor is changed by one unit, if all other variables are held constant. This seems easy to interpret, but most models have interdependencies between the predictors, such as interactions or polynomial terms, which means variables do not change independently. For more complex models, there may never be any direct interpretation of any of the model parameters. For example, in generalized additive models (Wood, 2006), knowing the form of the model and the parameter values does not give us insight into what the model says about the data."*. Parece ser coerente e deixam aqui um estrutura clara do que daqui em dinate irá ser feito e em que modos. Após isto os autores partem directamente para as ferramentas que irão ser usadas. Alegam que seria possível e fazivel usar gráficos e ferramentas comuns e ditas de estáticas (algo que será remtido e refutado na conclusão deste documento) mas preferem usar para dimensões elevadas técnicas interativas e que as duas ferramentes de eleição serão os grande tour (Wickham, 2011) e linked brushing. A primeira permite visualizar várias projeções dos dados em que se torna útil para obter uma visão ampla e geral do conjunto de dados sempre de forma dinâmica e a segunda, permite entender de forma fácil como um grupo ou conjunto de dados se relacionam sendo destacados no próprio gráfico.

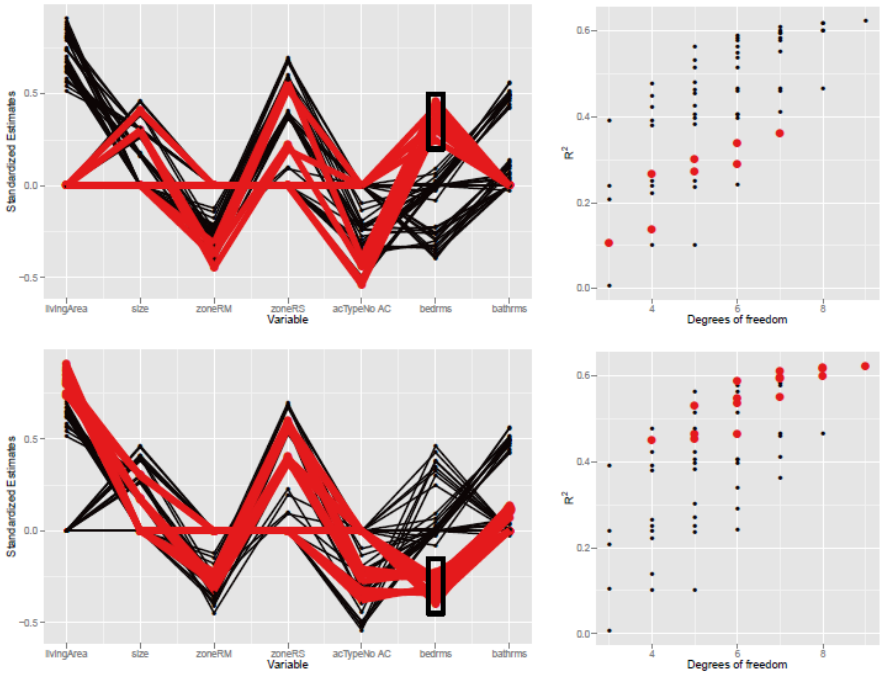
3. Casos de estudo revisitados

Nesta secção iremos revisitar alguns dos exemplos dados no artigo e tentar dissectar o que foi afirmado e abordado.Não revisitamos todos os exemplo pelo facto de em certa medida tornar-se repetitiva a exemplificação sendo que há particularidades nos vários model-vis, de grosso modo giram à volta do mesmo. O autores desde logo incluiem links de videos para aceder aos grand tours, possibilitado dessa forma ao leitor uma visão mais precisa e concisa do objetivo de cada exemplo dado. Os exemplos como já foi referido na introdução deste documento vão progredindo em termos de complexidade, se é que se pode dizer assim mas talvez mais por

adequação à estrategia em si que pretendem transmitir.

O caso de modelos de classficação da secção 3.2 é talvez o exemplo que mais deixa claro ao leitor do que está acontecer e certamente o mais interessante de todos. Muito devido aos videos associados ao exemplo que se fica com essa sensação concisa e explícita. https://vimeo.com/125405961, https://vimeo.com/125405962 . Por talvez ser o primeiro exemplo os autores entram em força e fazem um demonstração bastante conseguida. Mas é visivel ao longo do texto e dos exemplos, que de certa forma a atração visual e própria enfâse nestes vai dissipando-se de certo modo. Mais à frente entende-se o porquê desta afirmação.

O exemplo de modelos lineares com seis variáveis explicativas utilizando a base de dados NewHavenResidential, que demonstra o *model-vis* atraves da técnica *linked brushing* visivel pelas figuras 12 e 13. Os coeficientes standardizados de todos os modelos ajustados e os vários valores R^2 dos modelos são projetados em dois gráficos, respetivamente. Com o uso da técnica referida acima conectam todos os modelos numa só visualização. Passamos a demonstrar a figura 13 do artigo.



A razão pela qual é usada a figura em causa, é para ilustrar ao leitor que esta visualização fica longe de ser de fácil compreensão, nem de fácil percepção. Talvez até de choque, a própria escolha pelo vermelho tenha sido a menos apropriada. Gostos à parte, este conjunto de gráficos poderia ser debatido, que esta seria somente a projeção comum e clássica, que existiria uma versão em dimensão elevada (high dimensional space). Tal não sucede, deixando assim o leitor um pouco desiludido com o resultado. A visualização seja de dados e modelo, têm ou assumo

o paradigma da simplicidade visual por mais complexo que seja o modelo e os dados. O bom exemplo efetuado com os vários gráficos e videos na secção 3.2 é agora deitada por terra em prol de incutir ou até demonstrar uma estratégia elaborada pelos autores. De forma muito resumida as duas visualização (fig.12 e fig.13) são criadas para concluir que talvez possa existir colinearidades entre número de quartos (bedrms) numa casa e a dimensão em metros quadrados de uma casa (livingArea). Por mais compreensiva e ilustrativa estas visualizações sejam, existe aqui possibilidades em complicar a inferencia e selecção do modelo. Tal como se verifica com a própria visualização cirada de dificil intrepretação. Mesmo em casos em que a relação não seja tão óbvia, facilmente se verifica que as mesmas conclusões poderiam ser alcançadas através de um gráfico de dispersão usual e comumente utilizado. Até pelo coeficiente de correlação seria possivel obter suspeita e seguidamente investigar pelo gráfico de dispersão o que os autores concluiram, porém de forma muita mais fácil e rápida.

A secção 6 culmina com o exemplo mais complexo e completo. Os autores usam o grand tour e outros gráficos auxiliares para entender as caracteristicas do funcionamento de redes neuronais. O procedimento sugerido neste exemplo passa pela visualização de dados, modelos ajustados e model-vis, e termina com o entendimento do modelo. A abordagem tem um componente pedagógica muito interessante pois trata-se de uma rede somente de duas camadas, capacidade de visualizar os nós ocultos dentro do sistema da rede neuronal e demonstra a maneira como uma rede neuronal combina multiplas fronteiras logisticas para chegar a uma classificação não linear com um delimitação de fronteira bem clara. Porém, e face ao que foi feito em Tyner (2017), em que este mete à disposição dos leitores uma abordagem muito mais lúcida e clara. Oferece três opções, dependendo do grau de complexidade mas também das skills de cada praticante. Toma uma posição muito em linha do que é o ggplot2 e GGally, trabalhando em formato de extensão destes pacotes. Algo que não sucede no artigo em avaliação. Certo é, que Wickham e colegas afirma que o trabalho é uma continuidade de muito outros pacotes e autores, inclusive usam o pacote rggobi, do software GGobi e afirmam a facilidade que este pacote permitiu a realização do exemplo. Pelo facto de não haver acesso ao código usado e que o interface poder causar alguns problemas em certos utilizadores, fica pouco clara toda a implementação e futuros bugs que possam surgir. O Software em causa de acesso livre é sem dúvida útil, francamente bem conseguido e fornece uma série de análises especializadas que por norma são encontrados em softwares que acarretam alguns custos. Contudo, não deixa de ser um trabalho acrescido e de ambientação relativamente aos dois pacote referidos acima que estão bem estruturados e cimentados (ggplot e GGally). Nos

diversos videos ao longo do artigo mas principalmente o video associados a self-organizing map (https://vimeo.com/127615225), é possivel ver o grau de interação que se consegue ter com o software. Mas a própria falta de lucidez já referida deixa aquém do que se poderia esperar desta artigo em comparação com a secção 3.2. De notar que nesta secção 6, supostamente o apogeu do artigo, somente é fornecido um video de sete segundos (https://vimeo.com/767832) da iteração da rede neuronal em duas dimensões. Passa-se grande parte do artigo a falar em visualizar os modelos em dimensões superiores, por forma a conseguir entender toda a envolvência e interação existente e no momento em que talvez mais sentido faria em usar e abusar de grand tours e linked brushing tal não acontece. Em termos de método, encadeamento da explicação da rede neuronal, o uso sumários das 600 redes neuronais em termos visuais (fig 23), a figura 19 das iterações em tour (com interatividade em video em falta) foram mais ou menos conseguidas mas não passando das formas visuais já usuais e comuns a todos. Sendo a secção como um todo paupérrimo do que seria de esperar pelos os primeiros exemplo dados no artigo o leitor fica neste exemplo sempre grande interatividade ou até memso entusiasmo em visualizar um rede neuronal de forma dinâmica. Talvez pelo facto de este tipo de modelos necessitar uma grande quantidade de memória e velocidade tal não tenha sucedido no artigo. Porém nada é referido quanto a esse tipo de limitações na secção em si.

4. Conlusão

A tentativa de abarcar um grande conjunto de modelos num só artigo parece ser um tanto ambicioso e deixa azo a pouco esclarecimento sobre cada model-vis. Existe pouco aprofundamento de cada modelo, quase parecendo que o artigo em si se trata de um resumo de um outro documento de grande detalho e especificidade. A ideia subjacente é clara e os autores pretendem demonstrar que a estratégia elaborada por eles é aplicável aos diversos modelos. Em parte essa abordagem faz sentido, pois pode existir a vontade do que foi dito tornar-se eventualmente o gold standard na visualização. Em prol de um escalercimento e explicação detalhada teria sido interessante focar em dois ou três modelos e deixar ponto acento da qualidade e melhorias que estas estratégias podem trazer. Deu-se realce no inicio deste documento sobre as base de dados utilizados e voltamos a elas. Se a abordagem dos autores tivesse seguido com menos exemplos mas um aprofundamento de cada um, certamente que teria sido possivel efetuar a transição de dados nativos a pacotes do R para dados mais complexos, talvez até um caso real. Nessa situação seria para o leitor fácil de decifrar onde estão as grandes lacunas, como estas estartégias combatem as lacunas, o que se segue depois

de todo o método sugerido pelos autores e em que vertentes os trabalhos futuros poderiam eventualmente focar-se. Tyner (2017), é um exemplo claro que dirige seu foco em *Network* e diversas forma de as visualizar. Remetendo ferramentas ao leitor de como colmatar problemas e em que situações usar cada uma delas. Isto não sé palpável no artigo em questão e fica-se com uma mera directriz. Certo que deixa claro como um modelo pode ser visualizado, no espaço dos dados como parte de uma coleção de modelos mas como proceder à análise dessas visualizações fica sem clareza. O caso dos modelos lineares é evidente da falta de clareza. Sabido é, que um individuo ou analista por vezes depara-se com dificuldades em intrepretar ou tirar sentido de alguns gráficos mais comuns e dito de simples. Mostrar ou visualizar o máximo de dados possiveis possa em parte fazer sentido e dar vislumbres do que está a acontecer mas fica a questão de como um analista/pessoa irá realmente interagir com as visualizações sugeridas no artigo. As figuras 1, 6 ou 19 apresentam um visual bi-dimensional num espaço n-dimensional do dados, estes são bons exemplos de como irão ser estes gráficos compreendidos e intrepretados sem ser em formato tour ou interativo, e mesmo em caso visualizações em 3D ou superiores a própria análise é complexa e não sempre acessivel a todos. A critica do artigo ao longo deste texto tem sido pesada, é um facto mas muito derivado de se tratar dos autores que são pioneiros na visualização de dados em R. Tal com supramencionado, verifica-se uma decadência ao longo do texto não do conteudo em si mas sim das visualizações fornecidas. Outra questão que os autores descartam em sido demonstrado um prática eficaz na vertente de *high dimensional data* é a redução de dimensão tal como referido em Liu et al (2017), *"Due to the complexity of high-dimensional data, it is unlikely a single embedding (produced by dimension reduction) is sufficient for understanding every dataset. Instead, identifying multiple informative 2D projections automatically or semiautomatically is essential for exploring diferente aspects of the data. The subspace clustering methods either find clusters in subset of the dimensions (originated from data mining [58]) or cluster points that share a low-dimensional linear subspace (originated from machine learning [61]).These methods not only help in identifying multiple interesting projections but also address the challenges of the everincreasing complexity of the data (e.g., number of dimensions) by dividing them into lower dimensional subsets."*

Os autores deixam sem dúvida uma rampa de lançamento e porta aberta para explorar e inovar daqui em diante. Certo é que passado quase 6 anos da publicação do artigo, shiny tomou a frente no que toca a visualizações interativas e dinâmicas em R. Outros cientistas tal com Fernanda Viégas do MIT Media Lab tem feito progressos enormes na visualização de informação em larga escala. O trabalho realizado por Wickham e colegas deixou e levantou sem dúvida muitas

questões. As visualizações de modelos são cruciais e não deves ser negligenciados na análise de dados. Deve-se fomentar a procuar de técnicas e métodos que possam colmatar o handicap (velocidade, capacidade computacional ou capacidade de implementação) que exista em visualizar objetos ou dados em dimensões elevadas.

Referências

Gelman, A. & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press.

Genender-Feltheimer, A. (2018). *Visualizing High Dimensional and Big Data*. Procedia Computer Science. 140: 112-121, DOI: https://doi.org/10.1016/j.procs.2018.10.308.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. New York, USA: Springer

Johnson, R.A., Wichern D.W. (2007) *Applied multivariate statistical analysis*. (6th ed.). New Jersey, USA: Pearson Prentice-Hall

Liu, S., Maljovec, D., Wang, B., Bremer, P.T, Pascucci, V. (2017). *Visualizing High-Dimensional Data: Advances in the Past Decade*. IEEE Transactions on Visualization and Computer Graphics. 23 (3): 1249-1268, DOI: 10.1109/TVCG.2016.2640960

Tyner,S., Briatte, F., Hofmann, H. (2017). Network Visualization with ggplot2. The R Journal 9 (1): 27-59, DOI: https://doi.org/10.32614/RJ-2017-023

Wickham, H.(2015). *ggplot2: Elegante Graphics of Data Analysis*. New York, USA: Springer-Verlag.

Wickham, H., Cook, D., Hofmann, H., Buja, A. (2011) *tourr: An R package for exploring multivariate data with projections.* Journal of Statistical Software 40(2):1–18, DOI: 10.18637/jss.v040.i02

Wilkinson, L. (2005). *The Grammar of Graphics* (2nd. ed.). New York, USA: Springer