



7, 8 e 9
Março 2018
ÉVORA
Évora Hotel

GESTÃO DOS
RECURSOS HÍDRICOS:
NOVOS
DESAFIOS

USO DO RANDOM FOREST EM MODO EMBEDDED E WRAPPER PARA A SELEÇÃO DE ATRIBUTOS RELACIONADOS COM A POLUIÇÃO DA ÁGUA SUBTERRÂNEA POR NITRATOS

Maria Paula, MENDES¹; Victor, RODRIGUEZ-GALIANO²; Juan, LUQUE-ESPINAR³; Mario
CHICA-OLMO⁴

¹ Dr., CERIS-Investigação e Inovação em Engenharia Civil para a Sustentabilidade Lisboa 1049-001 - Portugal,
mpaulamendes@tecnico.ulisboa.pt

² Dr., Physical Geography and Regional Geographic Analysis, University of Seville, Sevilla 41004 - Espanha,
vrgaliano@us.es

³ Dr., Unidad del IGME en Granada, Granada 18006- Espanha, ja.luque@igme.es

⁴ Dr., Departamento de Geodinámica, Universidad de Granada, Avenida Fuentenueva s/n, Granada 18071- Espanha
mchica@ugr.es

Resumo

Para melhor combater a falta de qualidade das águas subterrâneas, devido à presença de nitratos em elevadas concentrações, é fundamental identificar as várias fontes de poluição. A seleção de atributos (SA) é um processo que apura um subconjunto dos atributos originais, de forma otimizada e de acordo com um certo critério. Vários métodos estatísticos podem ser utilizados na SA como wrappers e métodos embutidos (“embedded”). Este trabalho teve como finalidade a aferição do desempenho de um algoritmo de aprendizagem de máquina (“Machine Learning”), denominado “Random Forest” (RF) aplicado como “wrapper” e “embedded”, para a determinação dos atributos mais relevantes para predição da poluição da água subterrânea por nitratos. Para a indução do RF (embedded e wrapper) foram utilizadas 20 atributos (propriedades intrínsecas e potenciais fatores de poluição) conjuntamente com uma variável-alvo que consiste em concentrações de nitratos classificadas com base num valor de corte de 50 mg/l. Foram também adicionadas três variáveis de detecção remota extraídas de uma série temporal de índice de Vegetação da Diferença Normalizada (NDVI), para fornecer indicações da dinâmica do agroecossistema.

O RF wrapper com busca sequencial “forward floating” foi o método que obteve o melhor desempenho (erro de classificação=0.12). Este método apresentou uma boa interpretabilidade, onde três atributos foram selecionados: i) indústrias e instalações avaliadas de acordo com a sua capacidade de produção e total de emissões de azoto para a água num raio de 3 km, ii) pecuárias avaliadas pela sua produção de estrume num raio de 5 km e iii) NDVI acumulado após o mês dos valores máximos, sendo usado como representando a produtividade e rendimento das culturas agrícolas.

O estabelecimento de atributos que estão fortemente relacionados com a poluição por nitratos da água subterrânea pode contribuir para o estabelecimento de Programas de Ação, assegurando uma efetiva redução da poluição causada por nitratos e contribuindo para a sua prevenção.

Palavras-chave: Nitratos, água subterrânea, seleção de atributos, Random Forest, wrapper, seleção de atributos

Tema: Águas subterrâneas

1. INTRODUÇÃO

A poluição por nitratos nas águas subterrâneas é um dos problemas mais comuns por todo o mundo. A Diretiva Europeia dos Nitratos (91/271/EEC, 1991) foi desenhada com o intuito de se reduzir a poluição das águas causada ou induzida por nitratos de origem agrícola e impedir a propagação da poluição nas massas de água subterrânea ou superficial.

Ao contrário dos modelos com base nos processos físicos, os algoritmos de aprendizagem de máquina (AAM) focam-se na predição, baseando-se em propriedades conhecidas e aprendidas a partir dos dados de treinamento (Lary *et al.*, 2016). Estes métodos podem envolver algumas ou milhares de variáveis, permitindo o reconhecimento das relações entre atributos e uma variável-alvo (Kohavi and John, 1998). Os AAM englobam uma variedade de algoritmos, como por exemplo redes neuronais, “support vector machines”, algoritmos genéticos, árvores de decisão, “random forest”, classificadores “naive Bayes”, k-NN algoritmos, entre outros. Os algoritmos de aprendizagem de conjunto (“ensemble learning algorithms”) usam múltiplos AAM, alcançando um melhor desempenho, o qual não seria possível com um único AAM (Polikar, 2006). Dentro destes métodos, o “Random Forest” é uma combinação de árvores de decisão, em que para o crescimento de cada árvore é efectuada uma escolha aleatória (com reposição), a partir dos exemplos do conjunto de treinamento (Breiman, 2001). Atualmente têm surgido vários estudos em que o Random Forest é usado no estudo da poluição na água subterrânea causada por nitratos de origem agrícola (Nolan *et al.*, 2014; Ouedraogo *et al.*, 2017; Rodriguez-Galiano *et al.*, 2014, 2018; Tesoriero *et al.*, 2017; Wheeler *et al.*, 2015).

Comum a todos os métodos de AAM, inclusive ao “Random Forest”, está a possibilidade de os especialistas usarem todos os atributos existentes, ou reduzirem o seu número. A seleção de atributos é um processo que escolhe um subconjunto dos atributos originais, de forma otimizada e de acordo com um certo critério (Blum e Langley, 1997; Janecek *et al.*, 2008). O uso da SA tem como principais vantagens (i) melhorar o desempenho da previsão dos atributos; (ii) fornecer atributos mais rápidos e económicos e, (iii) contribuir para uma melhor compreensão do processo subjacente gerador dos dados. Vários métodos

estatísticos podem ser utilizados na SA como filtros, “wrappers” e métodos “embedded” (Guyon e Elisseeff, 2003).

Neste estudo vamos utilizar o “Random Forest” para a classificação em modo “wrapper” e “embedded” e, definir quais os atributos que melhor podem prever a probabilidade de

ocorrência de concentrações de nitratos na água subterrânea superiores ou iguais a 50 mg/l. Para tal serão identificadas quais das duas técnicas de SA tiveram melhor desempenho e, serão apontadas as fontes mais prováveis de poluição por nitratos na água subterrânea.

2. MATERIAIS E MÉTODOS

2.1 “Wrappers”

Os “wrappers” selecionam subconjuntos de atributos relevantes baseados na medida de desempenho do AAM. Estes algoritmos podem ser esquematizados em três fases: (i) a definição da medida de desempenho que serve como critério para a seleção do atributo e a estratégia de reamostragem para a validação; (ii) a seleção da estratégia de busca para o estabelecimento da ordem em que os subconjuntos são avaliados e, (iii) a adoção do algoritmo de aprendizagem. A medida do desempenho de predição do modelo final de classificação estabelecerá o subconjunto de atributos relevantes (Guyon e Elisseeff, 2003). Uma rotina de bootstrap pode ser incorporada ao “wrapper” para avaliação da generalização do modelo de predição. Diferentes estratégias de busca podem ser adotadas, tendo sido aplicado neste estudo uma busca sequencial que foi executada através de quatro maneiras distintas: seleção “backward” (SBS), seleção “forward” (SFS), seleção “forward floating” (SFFS) e seleção “backward floating” (SFBS). Uma descrição das diferentes estratégias pode ser encontrada em Reunanen, 2006. O método de bootstrap 0.632 + (Efron e Tibshirani, 1997) foi usado para a estimação do erro médio de classificação incorrecta (mmce) do “wrapper”. Este método utiliza pastas de teste para estimar o mmce, e assim a importância dos atributos.

2.2 Random Forest (RF) para a classificação

O valor de corte de 50 mg/l foi considerado como critério de distinção entre água poluída e não poluída de acordo com a Directiva dos Nitratos (91/676/EEC, 1991). Inicialmente, as nossas amostras de água foram classificadas em duas classes com base nas

concentrações de nitratos com valores superiores ou iguais a 50 mg/l (classificadas com o valor 1) e, concentrações abaixo desse valor de corte (classe igual a zero). As variáveis explicativas (ou atributos) e a variável resposta foram combinadas num conjunto de entrada de vetores das características. A variável binária de resposta foi usada para o treino do algoritmo.

O RF é um método que agrupa árvores de decisão (AD), crescendo cada uma das árvores a partir de um subgrupo do conjunto de treinamento, escolhido aleatoriamente e, com reposição (“bagging”) e, onde os atributos (variáveis preditivas) são escolhidas aleatoriamente (Breiman, 2001). Para o crescimento da AD, o algoritmo precisa de dividir o espaço amostral em regiões

cada vez mais homogêneas em relação à variável-alvo (poluídas ou não poluídas), de forma a que no final se possam obter amostras de água bem classificadas em cada região (Hastie *et al.*, 2009). O critério de divisão é baseado no índice de Gini, em que a selecção do atributo é baseada no valor mais baixo deste índice (Breiman *et al.*, 1984). No RF para a classificação cada árvore representa um voto numa classe e a classificação final baseia-se na maioria dos votos.

No RF “embedded” normalmente o subconjunto que não é seleccionado pelo “bagging” (denominado “out-of-bag”, oob) é utilizado para avaliação do desempenho e para a construção de uma medida da importância do atributo. A importância de um atributo também é medida com base na melhoria no critério da divisão num nó de uma árvore, e esta importância é acumulada para todas as árvores, para cada atributo.

O bootstrap 632+ foi usado para calcular a taxa de classificação incorrecta no modo “embedded” e no modo “wrapper”, sendo calculadas 20 florestas, permutando-se um atributo de cada vez.

2.3 Caso de estudo

O aquífero de Vega de Granada (VG) localiza-se numa planície aluvionar, na província de Granada no sul de Espanha. Este aquífero tem uma área aproximada de 200 km² (22 km × 8 km), com uma profundidade variável, podendo atingir, no centro, os 250 m e, diminuindo até aos 50 m para os limites a norte e a sul (Figura 1). O aquífero de Vega de Granada é constituído por depósitos Quaternários detriticos de granulometria diversa (i.e. gravilha,

areias, siltes e argilas), resultando num aquífero multicamada de origem sedimentar (Rodríguez-Fernández e de Galdeano, 2006). De acordo com Soldado (2009), os valores de transmissividade variam entre os 14 505 m²/dia e 63 m²/dia, situando-se os valores mais elevados nas áreas definidas pela confluência dos rios Monachil e Dilar para o rio Genil e na área ocidental, diminuindo esses valores para os bordos do aquífero.

Esta região é considerada semi-árida, com verões longos e secos (Maio a Setembro) e invernos chuvosos (Outubro a Abril). Os níveis piezométricos estão mais baixos entre

Agosto e Novembro e mais próximos da superfície entre Março e Maio (Castillo, 2005). A média anual da precipitação ronda os 450 mm, embora em alguns casos nas altitudes mais elevadas possa alcançar os 600 mm/ano (Luque-Espinar *et al.*, 2008). Esta área regista elevados valores de nitratos na água subterrânea, como consequência de décadas de uso de fertilizantes e, foi designada como zona vulnerável a nitratos (Comission, 2013).

2.4 Base de dados

A campanha de amostragem teve lugar em Novembro de 2016, após as culturas de verão e, foram recolhidas 110 amostras. Os estatísticos descritivos dos nitratos são os seguintes:

máximo de 547.3 mg/l, mínimo de 1.3 mg/l, primeiro quartil de 44.9 mg/l, terceiro quartil de 110.8 mg/l, uma média de 91.7 mg/l e uma mediana de 80.4 mg/l.

Vinte atributos foram selecionados para serem usados para a predição da probabilidade dos nitratos estarem acima dos 50 mg/l, na água subterrânea. Dentro destes atributos, temos variáveis intrínsecas do aquífero de VG (i.e. profundidade do nível freático, espessura da zona saturada, transmissividade e módulo do gradiente hidráulico), variáveis hidrológicas (i.e. direção do escoamento superficial e percentagem de aumento no percurso de descida mais íngreme de cada célula-“drop raster”), factores potenciais de contaminação por nitratos (i.e.. população existente baseada no census de Janeiro de 2014, distância dos centróides de população, ocupação do solo, distância a canais de irrigação e cemitérios, densidades de indústrias, instalações e agropecuárias) e, variáveis de deteção remota (i.e. NDVI- índice de Vegetação da Diferença Normalizada- máximo, tempo em que decorreu o máximo e valores do NDVI acumulado a partir do mês seguinte ao NDVI máximo).

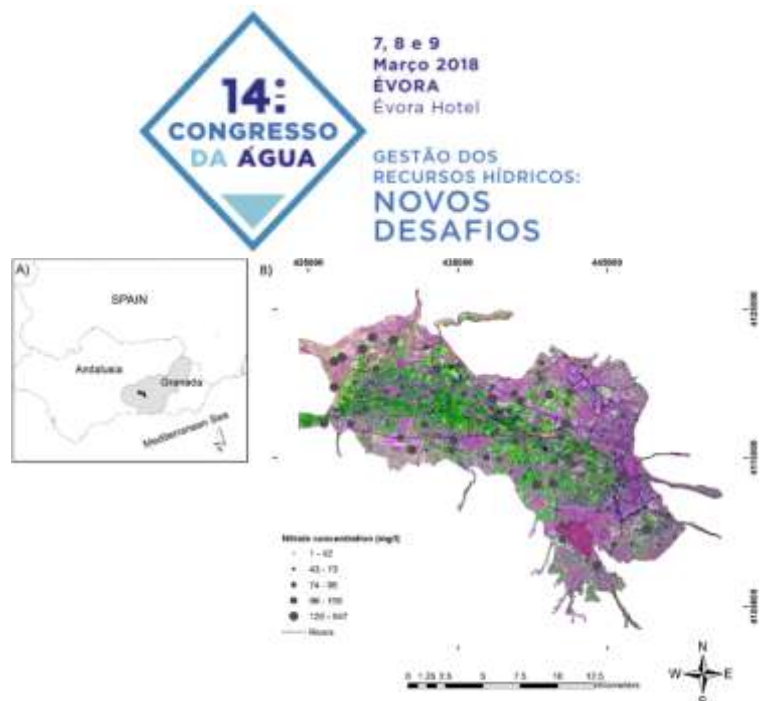


Figura 1. – A) Localização geográfica da área de estudo e B) distribuição dos pontos de amostragem com a respetiva classificação das concentrações de nitratos.

Todos os atributos foram obtidos em formato raster com uma resolução de 250 m, aplicando-se diferentes métodos de acordo com a natureza das variáveis. Deste modo, para rasterização de variáveis regionalizadas foi usada krigagem (Rodríguez-Galiano *et al.*, 2014). No caso das distâncias aos cemitérios e aos canais de irrigação optou-se por distâncias Euclidianas. Apenas foram considerados os canais de irrigação com problemas na qualidade da água, como resultado de descargas de efluentes (Luque-Espinar *et al.*, 2015).

A densidade de Kernel (Silverman, 1986) foi usada para criar novos atributos relacionados com potenciais fontes pontuais de poluição por nitratos, sendo aplicado uma média ponderada centrada na fonte. As indústrias (i.e. fábricas de fertilizantes e compostos azotados, preparação de produtos lácteos, fábricas de cerveja e, processamento e perservação de carne) e as instalações (e.g. ETARs e armazenamento e tratamento de resíduos não perigosos) foram classificadas consoante a sua capacidade de produção e total de emissões de compostos de azoto para a água em 2015 (Ministerio de Agricultura y Pesca, 2017).

Igualmente, as indústrias agropecuárias foram classificadas de acordo com o tipo e número de animais (Eurostat, 2013), considerando-se os coeficientes de excreção usados em Espanha (NIR, 2011). Para o cálculo das densidades das indústrias, instalações e agropecuárias foram adotados três raios de procura de 1000, 3000 e 5000 m.

No que concerne às fontes de poluição difusas, as categorias de ocupação do solo (legenda do nível III do mapa do Corine Land Cover de 2012) foram reclassificadas de acordo com o seu impacto potencial na poluição por nitratos (Ribeiro *et al.*, 2017).

A partir da série anual de 2016 do NDVI foram extraídas imagens compostas semanais com um tamanho de pixel de 250 m. O nível máximo de actividade fotossintética no coberto ($NDVI_{max}$), a altura do máximo da fotossíntese ($NDVI_{temp}$) e, o NDVI acumulado a partir do mês seguinte ao valor máximo ($NDVI_{pos-max}$) foram usados para indirectamente estimar a perda de azoto através da remoção pelas culturas e/ou, a lixiviação dos nitratos para a água

subterrânea devido a más práticas de adubação e/ou de gestão da rega. O $NDVI_{max}$ está associado ao tipo de vegetação, seu vigor e densidade, estando no nosso estudo, os valores mais elevados localizados nas áreas agroflorestais e os valores mais baixos nas áreas industriais e urbanas. O $NDVI_{temp}$ está dependente do tipo de vegetação e os valores mais elevados (Setembro e Outubro) estão localizados a noroeste e sudeste do aquífero de Vega de Granada. O $NDVI_{pos-max}$ pode ser usado como uma aproximação da produtividade da vegetação e rendimento da cultura estando os valores mais elevados localizados nas áreas agroflorestais e, seguidamente, nas áreas agrícolas de regadio. Os dois primeiros NDVIs podem indirectamente aferir as quantidades de fertilizantes usadas uma vez que relectem as diferentes necessidades das culturas em termos de azoto. O $NDVI_{pos-max}$ pode avaliar o N removido pelas culturas e a quantidade potencial de resíduos da colheita.

3. DISCUSSÃO DOS RESULTADOS

No caso do RF “embedded”, foi escolhido o modelo com o melhor compromisso entre o número de atributos necessários e o mmce, como base para a estimação da probabilidade de ocorrência de nitratos na água subterrânea. Apenas quatro atributos foram escolhidos como os mais importantes para a determinação das áreas de aquífero previsivelmente poluídas, sendo o mmce igual a 0.138 (Figura 2). Os atributos distância dos pólos urbanos e cemitérios, $NDVI_{max}$ e as agropecuárias a um raio de distância de 5 km definiram as áreas do VG com maior probabilidade de estarem poluídas.

A avaliação dos “wrappers” é baseada no desempenho do método de aprendizagem, onde o estabelecimento da ordem pelos quais os subgrupos de atributos são avaliados depende da estratégia de procura. No RF (SFFS) (“forward floating”), apenas três atributos foram seleccionados (densidade das indústrias e instalações a um raio de distância de 3000 m, densidade das instalações agropecuárias a um raio de 5000 m e, $NDVI_{pos-max}$).

O $NDVI_{pos-max}$ é uma aproximação para a produtividade e rendimento da cultura, e pode estar relacionado com um maior consumo de fertilizantes (EEA, 2015).

É igualmente interessante verificar que o erro obtido pelo RF (SFFS) (mmce = 0.120) é inferior ao obtido pelo RF “embedded” (Figura 3), sendo agora apenas escolhidos três atributos. Mesmo com um modelo em que foram usados todos os atributos, o RF teve um valor de mmce (0.135) superior a este wrapper. Contudo, os “wrappers” têm um custo computacional mais elevado quando comparados com o RF “embedded”.

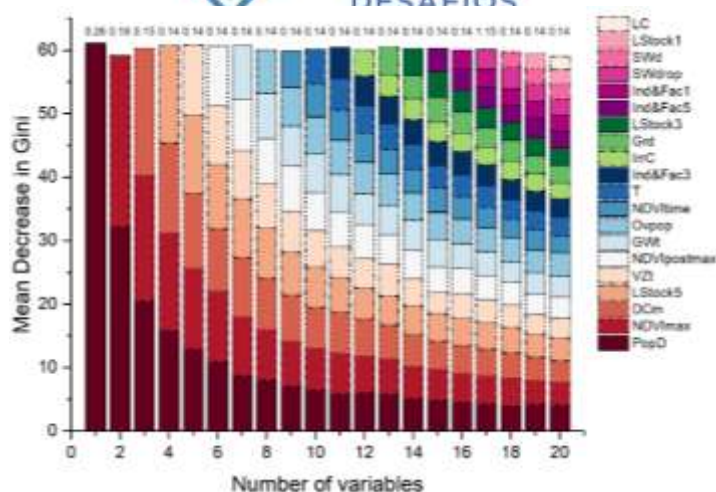
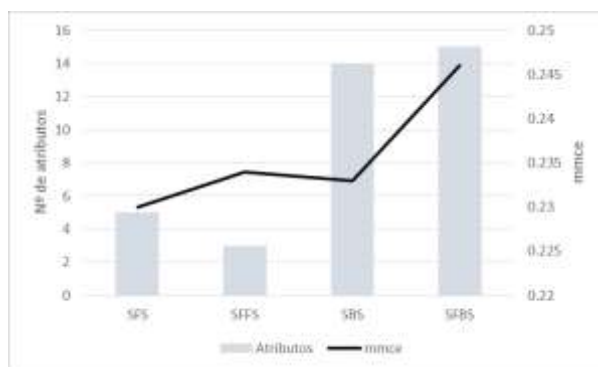


Figura 2. – Resultados do RF modo wrapper para as disrentes estratégias de busca (SFS,



SFFS, SBS e SFBS)

Figura 3. – Resultados do RF modo wrapper para as diferentes estratégias de busca (SFS, SFFS, SBS e SFBS)

Além disso, os resultados do RF (SFFS) estão em sintonia com o estudo de Pardo-Igúzquiza *et al.*, 2015, que apontam a agricultura irrigada e o sistema de esgotos da cidade de Granada como fontes de poluição por nitratos das águas subterrâneas no aquífero VG.

4. CONCLUSÃO

Como quase toda a área do aquífero de VG está ocupada por culturas de regadio, as séries temporais de NDVI provaram ser atributos importantes para a estimação da probabilidade de ocorrência de concentrações de nitratos indicadores de poluição. Tal deve-se ao facto de o $NDVI_{max}$ e o $NDVI_{temp}$ fornecerem informação relativa à fenologia das culturas, diferenciado deste modo, o tipo de cultura, e o $NDVI_{post-max}$ ser uma aproximação da biomassa da vegetação.

Embora computacionalmente mais exigente, o RF (SFFS) provou ser mais eficaz que o RF embedded. A SA através do RF embedded permitiu a selecção dos atributos mais importantes e a optimização do modelo de predição. Contudo, apenas com o uso de wrappers se conseguiu além dos dois objectivos atrás mencionados para a SA, a redução da dimensionalidade do espaço dos atributos.

De acordo com os resultados obtidos pelo o RF (SFFS), as principais fontes de poluição por nitratos são a agricultura, a agropecuária e, as agroindústrias. O NDVI_{max} e a sua relação com as áreas do VG com maior probabilidade de estarem poluídas, assinala a necessidade de se cumprirem boas práticas agrícolas. A actividade agropecuária é outro factor responsável pela poluição de nitratos na água subterrânea. Considerando o raio de influência de 5000 m, o espalhamento de lamas dessas explorações não deve estar a ser bem gerido. Perto das áreas urbanas (dentro de um raio influência de 3000 m), as águas residuais e/ou recolha de resíduos das cidades e agroindústrias podem não estar a receber o devido tratamento e, conseqüentemente, podem estar a contribuir para a poluição da água subterrânea por nitratos.

AGRADECIMENTOS

Maria Paula Mendes foi financiada por uma bolsa pós-doc FCT-MEC (SFRH/BDP/110346/2015). Os autores estão agradecidos pelo financiamento do Ministério da Economia, Indústria e Competitividade do Governo de Espanha (Projecto CGL2017-84739-R).

REFERÊNCIAS

Blum, A. L. and Langley, P. (1997) 'Selection of relevant features and examples in machine learning', *Artificial Intelligence*, 97(1), pp. 245–271. doi: [https://doi.org/10.1016/S0004-3702\(97\)00063-5](https://doi.org/10.1016/S0004-3702(97)00063-5).

Breiman, L. *et al.* (1984) *Classification and Regression Trees*. Taylor & Francis (The Wadsworth and Brooks-Cole statistics-probability series). Available at: <https://books.google.pt/books?id=JwQx-WOmSyQC>.

Breiman, L. (2001) 'Random Forests', *Machine Learning*, 45(1), pp. 5–32. doi: 10.1023/A:1010933404324.

Castillo, A. (2005). El acuífero de la Vega de Granada. Ayer y hoy (1966–2004). Agua, Minería y Medio Ambiente, Libro Homenaje al Profesor Rafael Fernández Rubio. López Geta *et al.*, pp. 161-172.

Comission E. on the Implementation of Council Directive 91/676/EEC Concerning the Protection of Waters Against Pollution Caused by Nitrates From Agricultural Sources Based on Member State Reports for the Period 2008–2011 Report From the Comission to the

Council and the European Parliament Brussels (2013), p. 11

91/271/EEC. Council Directive of 21.05.1991 concerning urban waste water treatment. Off. J. Eur. Communities (1991), p. 8

Efron, B. e Tibshirani, R. (1997) 'Improvements on Cross-Validation: The .632+ Bootstrap Method', *Journal of the American Statistical Association*. [American Statistical Association, Taylor & Francis, Ltd.], 92(438), pp. 548–560. Available at: <http://www.jstor.org/stable/2965703>.

Eurostat (2013). Nutrient Budgets – Methodology and Handbook. Eurostat and OECD, Luxembourg.

Guyon, I. e Elisseeff, A. (2003) 'An Introduction to Variable and Feature Selection', *J. Mach. Learn. Res.* JMLR.org, 3, pp. 1157–1182. Available at: <http://dl.acm.org/citation.cfm?id=944919.944968>.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) 'Random Forests', in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer New York, pp. 587–604. doi: 10.1007/978-0-387-84858-7_15.

Janecek, A. et al. (2008) 'On the Relationship Between Feature Selection and Classification Accuracy', in Saeys, Y. et al. (eds) *Proceedings of the Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery at ECML/PKDD 2008*. Antwerp, Belgium: PMLR (Proceedings of Machine Learning Research), pp. 90–105. Available at: <http://proceedings.mlr.press/v4/janecek08a.html>.

Khalil, A. et al. (2005) 'Applicability of statistical learning algorithms in groundwater quality modeling', *Water Resources Research*, 41(5), p. n/a–n/a. doi: 10.1029/2004WR003608.

Kohavi, R. e John, G. H. (1998) 'The Wrapper Approach', in Liu, H. and Motoda, H. (eds) *Feature Extraction, Construction and Selection: A Data Mining Perspective*. Boston, MA: Springer US, pp. 33–50. doi: 10.1007/978-1-4615-5725-8_3.

Lary, D. J. et al. (2016) 'Machine learning in geosciences and remote sensing', *Geoscience Frontiers*, 7(1), pp. 3–10. doi: <https://doi.org/10.1016/j.gsf.2015.07.003>.

Luque-Espinar, J. A. et al. (2008) 'Influence of climatological cycles on hydraulic heads across a Spanish aquifer', *Journal of Hydrology*, 354(1), pp. 33–52. doi: <https://doi.org/10.1016/j.jhydrol.2008.02.014>.

Luque-Espinar, J. A. et al. (2015) 'Seasonal occurrence and distribution of a group of ECs in the water resources of Granada city metropolitan areas (South of Spain): Pollution of raw drinking water', *Journal of Hydrology*, 531, pp. 612–625. doi: <https://doi.org/10.1016/j.jhydrol.2015.10.066>.

Ministerio de Agricultura y Pesca AyMA (2017). Registro Estatal de Emisiones y Fuentes Contaminantes. © PRTR España.

NIR (2011). Inventario de Emisiones de Gases de efecto Invernadero de España e Información adicional años 1990–2009. Comunicación a la Secretaría del Convenio Marco sobre el Cambio Climático y Protocolo de Kioto. Ministerio de Medio Ambiente, y Medio Rural y Marino Secretaría de Estado de Cambio Climático Dirección General de Calidad y Evaluación

Ambiental D.G., p. 706.

Nolan, B. T. et al. (2014) 'Modeling Nitrate at Domestic and Public-Supply Well Depths in the Central Valley, California', *Environmental Science & Technology*, 48(10), pp. 5643–5651. doi: 10.1021/es405452q.

- Nolan, B. T., Fienen, M. N. e Lorenz, D. L. (2015) 'A statistical learning framework for groundwater nitrate models of the Central Valley, California, USA', *Journal of Hydrology*, 531, pp. 902–911. doi: <https://doi.org/10.1016/j.jhydrol.2015.10.025>.
- Ouedraogo, I., Defourny, P. e Vanclooster, M. (2017) 'Validating a continental-scale groundwater diffuse pollution model using regional datasets', *Environmental Science and Pollution Research*. doi: 10.1007/s11356-017-0899-9.
- Pardo-Igúzquiza, E. *et al.* (2015) 'Compositional cokriging for mapping the probability risk of groundwater contamination by nitrates', *Science of The Total Environment*, 532, pp. 162–175. doi: <https://doi.org/10.1016/j.scitotenv.2015.06.004>.
- Polikar, R. (2006) 'Ensemble based systems in decision making', *IEEE Circuits and Systems Magazine*, 6(3), pp. 21–45. doi: 10.1109/MCAS.2006.1688199.
- Reunanen, J. (2006) 'Search Strategies', in Guyon, I. *et al.* (eds) *Feature Extraction: Foundations and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 119–136. doi: 10.1007/978-3-540-35488-8_5.
- Ribeiro, L., Pindo, J. C. and Dominguez-Granda, L. (2017) 'Assessment of groundwater vulnerability in the Daule aquifer, Ecuador, using the susceptibility index method', *Science of The Total Environment*, 574, pp. 1674–1683. doi: <https://doi.org/10.1016/j.scitotenv.2016.09.004>.
- Rodríguez-Fernández, J. e de Galdeano, C. (2006) 'Late orogenic intramontane basin development: the Granada basin, Betics (southern Spain)', *Basin Research*. Blackwell Science Ltd, 18(1), pp. 85–102. doi: 10.1111/j.1365-2117.2006.00284.x.
- Rodriguez-Galiano, V. *et al.* (2014) 'Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain)', *Science of the Total Environment*, 476–477. doi: 10.1016/j.scitotenv.2014.01.001.
- Rodriguez-Galiano, V. F. *et al.* (2018) 'Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods', *Science of the Total Environment*, 624. doi: 10.1016/j.scitotenv.2017.12.152.
- Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. Taylor & Francis (Chapman & Hall/CRC Monographs on Statistics & Applied Probability). Available at: <https://books.google.pt/books?id=e-xsrjsL7WkC>.
- Soldado, M.J. (2009). Metodología basada en SIG para el desarrollo de un sistema de un sistema soporte de decisión para la gestión de la calidad de los recursos hídricos subterráneos de la "Vega de Granada". Universidad de Granada [357 pp.].
- Tesoriero, A. J. *et al.* (2017) 'Predicting redox-sensitive contaminant concentrations in groundwater using random forest classification', *Water Resources Research*, 53(8), pp. 7316–7331. doi: 10.1002/2016WR020197.
- Wheeler, D. C. *et al.* (2015) 'Modeling groundwater nitrate concentrations in private wells in Iowa', *Science of The Total Environment*, 536, pp. 481–488. doi: <https://doi.org/10.1016/j.scitotenv.2015.07.080>.