

MODELOS DE INTELIGÊNCIA ARTIFICIAL NA PREDIÇÃO DOS EPISÓDIOS DE INTERNAMENTO POR PNEUMONIA EM PORTUGAL CONTINENTAL

ROCHA¹, Jorge; ROQUETTE², Rita; VIANA³, Cláudia Morais; DOROSH⁴, Valeriya

¹ Centro de Estudos Geográficos e Laboratório Associado TERRA, Instituto de Geografia e Ordenamento do Território, Universidade de Lisboa, Portugal, jorge.rocha@campus.ul.pt

² NOVA IMS, NOVA Universidade de Lisboa, Portugal, rita.roquete71@gmail.com

³ Centro de Estudos Geográficos e Laboratório Associado TERRA, Instituto de Geografia e Ordenamento do Território, Universidade de Lisboa, Portugal, claudiaviana@campus.ul.pt

⁴ Centro de Estudos Geográficos, Instituto de Geografia e Ordenamento do Território, Universidade de Lisboa, Portugal, valeriyadorosh@campus.ul.pt

RESUMO: As doenças respiratórias são a terceira causa de morte em Portugal, ficando apenas atrás das doenças cardiovasculares e do cancro. A pneumonia, em particular, tem um peso elevado quando se fala das doenças respiratórias. Trata-se de uma doença com elevado número de internamentos e de óbitos, em particular nas faixas etárias mais avançadas e vulneráveis nomeadamente a partir dos 65 anos. Existe assim a necessidade de compreender a distribuição espacial deste fenómeno, e é aqui que a geografia e em particular os sistemas de informação geográfica ligados à área da saúde podem dar o seu maior contributo. OS modelos usados permitiram identificar os possíveis fatores que influenciam a ocorrência desta doença, nomeadamente as condições de vida e de habitação das famílias portuguesas, a qualidade e quantidade dos serviços de saúde que prestam auxílio à população, ou mesmo as condições ambientais e a emissão de poluentes.

PALAVRAS-CHAVE: Inteligência Artificial; modelos preditivos; pneumonia; assimetrias regionais; Portugal Continental

1. INTRODUÇÃO

A evolução do panorama das doenças respiratórias, em Portugal e no mundo, tem colocado desafios crescentes aos sistemas de saúde, relacionados com uma mudança de paradigma epidemiológico caracterizada por um aumento crescente das doenças respiratórias crónicas (DRC) em oposição a um decréscimo das doenças respiratórias agudas (Froes *et al.*, 2013).

O aumento do peso das doenças respiratórias está relacionado não só com os efeitos a curto e longo prazo do tabagismo, mas também com o aumento progressivo da esperança de vida e das condições ambientais e de vida das populações, levando a que as doenças respiratórias se tornem numa importante causa de morbilidade e,

mortalidade em faixas etárias mais avançadas, nomeadamente acima dos 65 anos (Ruhnke *et al.*, 2011).

A nível nacional, existem assimetrias regionais para a taxa padronizada de mortalidade por doença respiratória, destacando-se as regiões autónomas da Madeira e Açores pelos seus valores elevados e a região de Lisboa e Vale do Tejo (LVT) pela positiva. No que se refere a internamentos, as doenças respiratórias constituem a quinta principal causa de internamento e a primeira causa de mortalidade intra-hospitalar (Bárbara *et al.*, 2017).

Deste modo pretende-se caracterizar a população internada por pneumonia, por sexo e idade; analisar a distribuição mensal dos internamentos e óbitos por pneumonia; analisar a distribuição e evolução espacial dos internamentos por pneumonia em Portugal Continental a fim de identificar desigualdades regionais; identificar os fatores que influenciam a localização geográfica dos episódios de internamento, de modo a identificar áreas de concentração do fenómeno.

Um dos desafios da modelação de fenómenos espaciais é a escolha de um modelo matemático que seja capaz de se adequar ao problema estudado, um modelo que seja capaz de caracterizar as dinâmicas e os comportamentos dos dados de *input* e *output*. Neste contexto, foram testadas duas estruturas de redes com diferentes parâmetros de organização, nomeadamente redes perceptrão multicamada (MLP) e de função de base radial (RBF).

A nível dos Centros Hospitalares / Unidades Locais de Saúde, é no Centro Hospitalar Universitário de Coimbra e nos Centros Hospitalares de Lisboa Norte e Lisboa Centro que os episódios de internamento e de óbitos são maiores. Se olharmos antes para os fatores explicativos que identificam possíveis causas da doença e resultante internamento, temos as condições de vida e de habitação, com muitos modelos a darem muita importância a estes fatores, onde ainda se destacam as condições dos serviços de saúde com o número de consultas e de médicos ou a nível ambiental o nível de emissões de poluentes como principais causadores da doença.

Destaca-se uma maior concentração de episódios de internamento no centro e norte de Portugal Continental, com os municípios de Coimbra, Pombal, Leiria e Lisboa a apresentarem o maior número de episódios de internamento segundo os modelos de

redes neuronais (MLP e RBF) utilizados.

A utilização de modelos preditivos como as redes neuronais é muito útil para o estudo e compreensão de fenómenos espaciais através de modelação espacial, em particular na área da saúde com estudo das mais diversas doenças não só no caso da pneumonia, mas aquelas que mais afetam a população portuguesa. Sendo que através destes métodos é possível identificar áreas de concentração de casos e identificação de fatores de risco, contribuindo assim para uma melhor gestão dos recursos disponíveis na prevenção e mitigação de casos de pneumonia ou de outras doenças.

2. METODOLOGIA

A área de estudo para a realização deste estudo corresponde a Portugal Continental. Portugal Continental apresenta uma área de 89 101km², divididos em 18 distritos, 278 municípios e 4050 freguesias (2011) / 2882 freguesias (2016). Segundo o Instituto Nacional de Estatística (INE) e com base nos Censos de 2011 existem em Portugal Continental 10 047 621 habitantes em Portugal Continental.

Foram utilizados dados referentes ao número de internamentos por pneumonia durante o período 2000-2016, fornecidos pela Direção Geral de Saúde.

Os municípios com maiores taxas de internamento estão localizados no centro do país. O município com maior taxa de internamentos é o de Vila de Rei com 208,57 internamentos por cada 10 000 habitantes, seguido de Macedo de Cavaleiros no norte do país com uma taxa de 172,09 internamentos e pelo município de Pampilhosa da Serra com 153,98 internamentos. Por outro lado, os municípios que apresentam as taxas de internamento mais baixas são os da Trofa com 23,28, o de Vizela com 30,88 e o de Paços de Ferreira com 31,11.

É também no centro do país que se verificam as maiores taxas de internamento entre a população com 65 ou mais anos, com destaque para Macedo de Cavaleiros com 486,35, Seia com 479,46 e Penela com 466,36. Já os municípios com as menores taxas encontram-se no norte do país, sendo eles Vila Nova de Foz Côa com 87,87, Melgaço com 88,86 e Pinhel com 93,66.

No caso em estudo (número total de internamentos por pneumonia por município para o ano de 2011) foram testados vários modelos MLP (quadro I), com o

Desigualdades em saúde, desigualdades no território:
desafios para os países de língua portuguesa em contexto de pós pandemia

intuito de encontrar aquele que apresenta melhores resultados. Em todos os testes mantiveram-se as 63 variáveis independentes e a variável dependente (Número de internamentos de 2011), sendo alterados outros parâmetros tais como o número de camadas ocultas e as funções de ativação de entrada e de saída.

Quadro I. Estrutura e desempenho dos modelos MLP.

Número do Modelo	Número de camadas ocultas	Número de Unidades na Camada Oculta 1 ^a	Número de Unidades na Camada Oculta 2 ^a	Função de ativação	Função de ativação saída	Treino		Teste	
						Soma dos erros quadráticos	Erro relativo	Soma dos erros quadráticos	Erro relativo
1	1	12	0	Tangente hiperbólica	Identidade	1,282	0,078	1,872	1,172
2	1	12	0	Curva Sigmoide	Identidade	5,997	0,387	0,482	0,210
3	1	12	0	Curva Sigmoide	Tangente hiperbólica	0,015	0,007	0,017	0,164
4	1	12	0	Curva Sigmoide	Curva Sigmoide	0,001	0,001	2,479	0,606
5	2	12	9	Tangente hiperbólica	Identidade	0,872	0,053	60,366	0,687
6	2	12	9	Curva Sigmoide	Identidade	5,720	0,369	0,528	0,265
7	2	12	9	Curva Sigmoide	Tangente hiperbólica	0,688	0,363	0,045	0,219
8	2	12	9	Curva Sigmoide	Curva Sigmoide	,244	0,266	5,723	0,824

A rede neuronal RBF tal como a MLP cria um modelo preditivo para um ou mais variáveis dependentes, que no caso estudo são os episódios de internamento para o ano de 2011, que com a ajuda de variáveis independentes vai criar o modelo. As redes RBF são compostas por dois tipos normalizadas ou ordinárias, sendo que pode utilizar como funções de ativação Softmax e Exponencial e como função de ativação de saída identidade.

Em todos os modelos RBF testados (quadro II), com 63 variáveis independentes e a variável dependente (Número de internamentos de 2011), foram alterados parâmetros tais como a função RBF, o número de camadas ocultas e as funções de ativação de entrada e de saída.

Quadro II. Estrutura e desempenho dos modelos RBF

Número do Modelo	Número de camadas ocultas	Função de ativação	Função RBF	Função de ativação saída	Treino		Teste	
					Soma dos erros quadráticos	Erro relativo	Soma dos erros quadráticos	Erro relativo
1	6				3,466	0,210	61,296	0,780
2	5	Softmax	Normalized	Identidade	0,598	0,043	0,988	0,607
3	4				1,545	0,088	0,942	0,812
4	6	Exponencial	Ordinary	identidade	0,431	0,032	1,869	1,232
5	8	Exponencial	Ordinary	identidade	0,772	0,048	1,450	0,960

3. RESULTADOS

Na construção dos modelos MLP com uma camada oculta foram testadas três funções de ativação (curva sigmoide, tangente hiperbólica e identidade), bem como vários números de unidades nas camadas ocultas.

As arquiteturas da rede que foram testadas e que apresentaram os melhores resultados prendem-se com a função de ativação de saída e o número de camadas ocultas. Em relação aos restantes parâmetros dos modelos continuaram iguais, sendo que o número de camadas ocultas foi de 1, a função de ativação foi a curva sigmoide, nas unidades de saída foram utilizadas 12 unidades na 1ª camada oculta.

Em suma, entre os modelos MLP com 1 camada oculta e 12 unidades na 1ª camada oculta o que apresentou melhores resultados foi o modelo três. Podemos verificar que o erro relativo de treino (0,015) é inferior ao erro relativo de teste (0,164). Em comparação com o modelo dois notamos que o erro relativo do treino (0,387) é superior ao erro relativo do teste (0,210).

Se analisarmos os modelos dois e três em termos de importância relativa e normalizada – capacidade que cada variável independente tem em explicar o número de internamentos em 2011 – verificamos que o modelo três foi o que apresentou melhores resultados relativamente ao erro relativo (0,164), dando mais importância a variáveis relativas ao número de famílias em determinado tipo de alojamento e sendo a variável mais influente o número farmácias e de postos farmacêuticos móveis com 0,106 de importância relativa, seguindo-se o número de famílias em alojamentos com ar condicionado com 0,065 e do número de famílias em alojamentos sem retrete com 0,045. Ainda entre as variáveis com mais importância para o modelo encontra-se o

número de famílias em alojamentos construídos entre 1946-1960 com 0,035 de importância. Pelo contrário, as variáveis com menos importância para o número de internamentos em 2011 prendem-se o número de emissões de poluentes nomeadamente o mercúrio e as partículas com diâmetro inferior a 10 µm com 0,005 de importância relativa e o número de famílias em alojamentos construídos entre 1991-1995 com 0,004 de importância. Para ajudar a explicar os resultados do modelo pode, que representa os resíduos vs o valor predito para o modelo três.

O modelo dois dá mais importância às variáveis relacionadas com a idade dos edifícios e às emissões de poluentes. As variáveis que explicam melhor o número de internamentos em 2011 para este modelo são o número de famílias em alojamentos construídos entre 1919-1945 com 0,060 e até 1919 com 0,053 de importância relativa, seguindo-se as emissões de óxido de nitrogênio com 0,037. Destacar ainda o número de médicos especialistas em medicina geral familiar com 0,036. No final, surgem como variáveis menos explicativas para o número de internamentos em 2011 o número de médicos especialistas em estomatologia (0,003), as emissões de cádmio (0,004) e o número de alojamentos só com retrete e aquecimento com (0,005).

Na construção dos modelos com duas camadas ocultas, foram testadas as mesmas três funções de ativação (curva sigmoide, tangente hiperbólica e identidade) que foram utilizadas nos modelos com uma camada oculta, com as mesmas 12 unidades na 1ª camada oculta e com nove unidades na 2ª camada oculta.

Em relação às arquiteturas da rede que foram testadas, as que apresentaram os melhores resultados com duas camadas ocultas foram as que usaram as funções de ativação curva sigmoide e a função de ativação de saída identidade (0,265 de erro relativo) e tangente hiperbólica (0,219 de erro relativo).

Resumindo os modelos MLP com duas camadas ocultas, 12 unidades na 1ª camada oculta e nove unidades na 2ª camada oculta, o que apresentou melhores resultados foi o modelo sete. Podemos verificar que o erro relativo de treino (0,363) é superior ao erro relativo de teste (0,219). Em comparação com o modelo seis verificamos que o erro relativo no treino (0,369) é superior ao erro relativo de teste (0,265).

No modelos seis e sete em termos de importância relativa e normalizada que os

modelos atribuem às das variáveis independentes observamos que o modelo sete, o que apresentou os melhores resultados (0,219 de erro relativo) entre os modelos testados com duas camadas ocultas, vemos que este modelo dá mais importância a variáveis relacionadas com os cuidados hospitalares e a idade dos edifícios onde as famílias moram. A variável que mais importância tem para o modelo sete é o número de camas hospitalares com 0,050 de importância relativa, seguido do número de farmácias e postos farmacêuticos móveis com 0,047 e do número de famílias em alojamentos construídos entre 1946-1960 (0,042) e dos edifícios construídos entre 1961-1970 (0,041). O modelo dá ainda importância ao número de médicos (0,038) e às emissões do poluente óxido de nitrogênio (0,032). Entre as variáveis a que o modelo dá menos importância estão o rácio de população residente com >65 anos (0,003), o número de famílias em alojamentos sem água canalizada (0,004) ou o rácio de habitantes por farmácias e postos farmacêuticos móveis (0,004).

No modelo seis, as variáveis que se identificam como sendo as mais relevantes para o número de internamentos têm que ver com as condições dos alojamentos das habitações, com o número de famílias em alojamentos se nenhum sistema de aquecimento a ser a mais importante com 0,051 de importância, seguida do número de famílias em alojamentos com banho com 0,047 e o número de famílias em alojamentos só com água com 0,035. Do outro lado, as que menos relevância têm para os internamentos segundo este modelo são o número de famílias em alojamentos sem qualquer instalação (0,004), a percentagem de qualidade da água para consumo humano (0,005) ou o rácio de habitantes por médicos hospitalares (0,006). Para complementar a informação acima analisada, que representa os resíduos vs o valor predito para o modelo seis.

Depois de realizados os modelos RBF, o modelo que apresentou melhores resultados foi o modelo dois com 0,607 de erro relativo. Na construção deste modelo foi utilizada a função RBF Normalized, com cinco camadas ocultas, e a função de ativação foi *Softmax*, e a função de ativação de saída foi Identidade. O segundo modelo com melhores resultados foi o modelo um com 0,780 de erro relativo, este modelo foi composto pela mesma arquitetura do modelo dois, diferenciando-se apenas na utilização de seis camadas ocultas ao invés de cinco.

A importância relativa e normalizada para o modelo RBF dois aponta para as condições dos alojamentos e o número de consultas médicas como as variáveis com mais importância, onde o número de famílias em alojamentos com recuperador de calor (0,031) de importância relativa é o mais destacado. Seguem-se o número de consultas em centros de saúde de pediatria com 0,029 e o número de famílias em alojamentos construídos entre 1981-1990. Do lado oposto, as variáveis a que o modelo dá menos importância são o número de famílias em alojamentos só com aquecimento (0,009), seguindo-se o rácio de habitantes por farmácias e postos farmacêuticos móveis e a qualidade da água para consumo humano com 0,009 e 0,010 respetivamente.

4. CONCLUSÕES

Com as questões relacionadas com a saúde na ordem do dia, existe uma consciencialização da importância que é estudar temas de saúde. A geografia e os SIG são uma importante parte quando é necessário o desenvolvimento de modelos que permitam aos decisores e planeadores uma maior compreensão da distribuição espacial dos fenómenos no território, de forma a apoiar a tomada de decisões e dar respostas de forma quase imediata aos problemas territoriais, a fim de melhorar a qualidade de vida das populações.

Sabendo que as doenças respiratórias são a terceira causa de morte em Portugal, ficando apenas atrás das doenças cardiovasculares e do cancro, o estudo da distribuição espacial da pneumonia é de grande importância, uma vez que se trata da patologia com o maior número de internamento e de óbitos entre tal causa de morte. Como seria de esperar, é nas faixas etárias mais avançadas (≥ 65) que ocorre o maior número de internamentos e de óbitos, com esta a representar sensivelmente 70% do total de internamentos e 90% dos óbitos totais. A idade média dos doentes internados ronda os 66 anos e a idade média dos pacientes falecidos os 80 anos.

A distribuição espacial dos episódios de internamento e dos óbitos, encontram-se distribuídos pelo território de uma forma homogénea. Apenas a destacar uma maior concentração de episódios de internamento no centro e norte de Portugal Continental, com os municípios de Coimbra, Pombal, Leiria e Lisboa a apresentarem o maior número de episódios de internamento segundo os modelos de redes neuronais

(MLP e RBF) utilizados.

A nível dos Centros Hospitalares / Unidades Locais de Saúde, é no Centro Hospitalar Universitário de Coimbra e nos Centros Hospitalares de Lisboa Norte e Lisboa Centro que os episódios de internamento e de óbitos são maiores. Se olharmos antes para os fatores explicativos que identificam possíveis causas da doença e resultante internamento, temos as condições de vida e de habitação, com vários modelos a darem muita importância a estes fatores, onde ainda se destacam as condições dos serviços de saúde com o número de consultas e de médicos ou a nível ambiental o nível de emissões de poluentes como principais causadores da doença.

A utilização de modelos preditivos, como as redes neuronais, é muito útil para o estudo e compreensão de fenómenos espaciais através de modelação espacial, em particular na área da saúde para estudo das mais diversas doenças não só no caso da pneumonia, mas aquelas que mais afetam a população portuguesa. Através destes métodos é possível identificar áreas de concentração de casos e identificação de fatores de risco, contribuindo assim para uma melhor gestão dos recursos disponíveis na prevenção e mitigação de casos de pneumonia ou de outras doenças.

BIBLIOGRAFIA

- Froes F, Diniz A, Mesquita M, Serrado M, Nunes B. (2013). Hospital admissions of adults with community-acquired pneumonia in Portugal between 2000 and 2009. *European Respiratory Journal*, 41(5), 1141-6. <https://doi.org/10.1183/09031936.00216711>
- Ruhnke, G. W., Coca-Perraillon, M., Kitch, B. T., Cutler, D.M. (2011). *Marked reduction in 30-day mortality among elderly patients with community-acquired pneumonia. The American Journal of Medicine*, 124(2), 171-178. <https://doi.org/10.1016/j.amjmed.2010.08.019>
- Welte T, Torres A, Nathwani D. (2012). Clinical and economic burden of community-acquired pneumonia among adults in Europe. *Thorax*, 67(1), 71-79. <https://doi.org/10.1136/thx.2009.129502>