Boletim

SPE
Sociedade Portuguesa de Estatística

Publicação semestral *primavera de 2020*

INE - 85 anos de estatísticas a servir o país

Mensagem do Presidente do Instituto Nacional de Estatística
Francisco Lima 8

Inovação nas Estatísticas Oficiais - Principais desafios do INE
Instituto Nacional de Estatística 10

Modelos de regressão para dados de contagem: uma aplicação ao Inquérito ao Transporte Rodoviário de Mercadorias
Inês Rodrigues 14

Short-Term Regional Demographic Forecasts with Time Series Methods and Machine Leaming Algorithms
Jorge M Bravo e Edviges Coelho 20

Support Vector Machine para Imputação e Edição de Valores - O caso das Declarações Mensais de Remuneração das Empresas
Filipe Santos e Pedro Campos 30

O papel da intermediação do setor financeiro - quando o todo é menor do que a soma das partes
Filipa Lima, Sónia Mota e Ângela Coelho 36

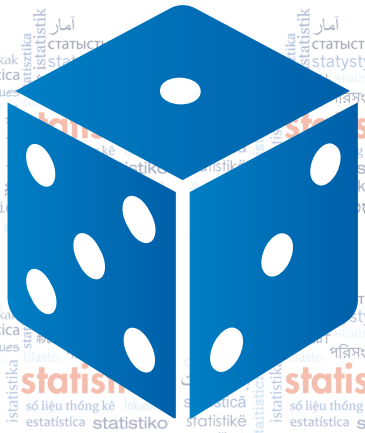
O INE como (pro)motor da Literacia Estatística
Francisco Correia 42

INE: preservar a memória institucional
Paula Marques 46

Editorial 1
Mensagem da Presidente 3
Notícias 4
Enigmística 7
INE - 85 anos de estatísticas a servir o país 8
Ciência Estatística 57
Edições SPE Mini-Cursos 59
Prémios SPE 60

Informação Editorial

Endereço: Sociedade Portuguesa de Estatística.
Campo Grande. Bloco C6. Piso 4.
1749-016 Lisboa. Portugal.
Telefone: +351.217500120
e-mail: *spe*@spestatistica.pt
URL: https://www.spestatistica.pt
ISSN: 1646-5903
Depósito Legal: 249102/06
Tiragem: Edição digital
Execução Gráfica e Impressão: Gráfica Sobreirense
Editor: Fernando Rosado, *fernando.rosado@fc.ul.pt*



SPE

Sociedade Portuguesa de Estatística

https://www.spestatistica.pt

Editorial

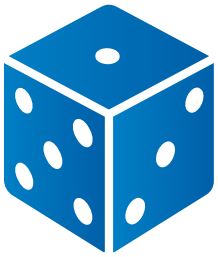
… e, nos 40 anos da SPE, nasceu… Um Boletim Outlier!

A Estatística alimenta-se e vive de observações, digamos: os dados.
No início, toda essa informação, especialmente os novos, cada vez mais importantes, "*big data*", têm o mesmo valor. O estudo, no entanto, permite decidir sobre as diferenças entre aquelas observações.
E, é assim que, alguma (ou algumas) se salientam como distintas. Explicar: Porquê? É um dos grandes desafios. Este é (também) o caminho simples da pesquisa de *outlier(s)*.
Para a presente edição do *Boletim SPE*, os seus criadores, editor, coeditores e autores tudo planearam, como sempre acontece, para a sua construção. Tudo seguia o caminho habitual, quando chegaram "os novos tempos" que vivemos – tão avassaladores quanto desafiantes.
Surgiram então as mais diversas dificuldades logísticas que inviabilizaram a pontualidade sempre desejada como característica do *Boletim SPE*. O tempo e as agendas mostraram grandes novidades. E tudo isso originou um grande atraso na publicação deste *Boletim primavera 2020*. Também, o Congresso SPE em data simbólica foi mexido, como se notícia. Toda a programação aniversária fica "histórica".
A vivência pandémica dos dias de hoje: Tudo alterou! O ritmo científico e a própria Ciência Estatística enfrentam novos grandes desafios. Esta é uma grande vantagem (também) para a Estatística; pois a resposta à chamada de intervenção e ação aos mais diversos níveis permite-lhe afirmar-se, ainda mais, como Ciência fundamental na decisão e como tal (ainda mais) reconhecida.

O *Boletim outono de 2019*, como sabemos, teve o privilégio de iniciar uma colaboração periódica e mais regular por parte do Instituto Nacional de Estatística – INE. A secção *SPE e a Comunidade* ficou assim mais rica.
Neste Boletim, temos um novo enorme prazer ao poder receber o INE e os seus convidados que nas nossas páginas transmitem informação, conhecimento e história. Ao Prof. Pedro Campos e ao Dr. Carlos Marcelo, neste número os coeditores e, acima de tudo, aos autores são devidas as maiores felicitações e agradecimentos.
A SPE através do *Boletim primavera de* 2020 também se sente honrada ao fazer parte integrante das comemorações aniversárias que neste ano decorrem no *Instituto Nacional de Estatística.*
A SPE e o INE desde sempre estiveram muito próximos na partilha dos mais diversos acontecimentos estatísticos nacionais e internacionais. São inúmeras as ações em que a SPE teve o apoio do INE, fundamentais para a sua concretização. Em 2005, por ocasião da edição do livro *Memorial da SPE*, na celebração dos 25 anos, tivemos a participação de dois antigos Presidentes, Manuel José Vilares e Paulo Gomes que contribuíram, a páginas 125 e 163, com textos sobre o Sistema Estatístico Nacional.
Como testemunho pessoal, porque tive a felicidade de viver intensamente esse momento enorme de afirmação da Estatística e dos estatísticos portugueses no panorama da comunidade internacional; permito-me incluir no Memorial dos 85 anos do INE que aqui perpetuamos: a realização em Lisboa, do *ISI* 2007, do *International Statistical Institute - ISI* levado a cabo por uma equipa científica liderada pela Prof. Ivette Gomes e uma executiva comandada pelo Prof. Paulo Gomes. Foi um marco nacional e internacional onde a SPE e o INE inscreveram Portugal na lista de referência do *ISI*.

Pelo exposto, este Boletim é histórico. E como tal deve ser inserido no memorial do ano aniversário que a "comunidade estatística" vive. O *INE* com a experiência e o saber dos 85 e a *SPE* na vivência dos quarenta souberam construir uma aliança de que todos nos devemos orgulhar. Conseguimos construir um legado que, seguramente, as gerações futuras dos estatísticos portugueses saberão fazer render e aumentar.
Assim, numa análise estatística que todos entendem, este Boletim pode (também) ser eleito como um *outlier* – uma observação (edição) especial! Mas, como em todo o estudo de *outliers*, de imediato surge a problemática, quando tudo está dito e feito, mantendo-se sempre a questão inicial: O que é um *outlier* e como deve ser estudado. Garantida fica (pelo menos) a satisfação pelo dinamismo da Estatística portuguesa que testemunhamos e que é uma concretização do primeiro objetivo estatutário da SPE.

Este *Boletim SPE primavera de 2020* inclui o novo logotipo da Sociedade Portuguesa de Estatística. Em época aniversária, esta é uma ação, também simbólica, que a Direção decidiu agregar às iniciativas da década festiva que se comemora. Este é o terceiro logotipo da SPE.



SPE
Sociedade Portuguesa
de Estatística

Um "simbólico dado" continua como bandeira da SPE – no início com a sigla (5 pintas) SPEIO ao alto, em 1991 "reequilibrado" com SPE; e aos 40 anos de vida, com uma nova perspetiva "do acaso". *Alea jacta est!* Os dados estão lançados! Mas, não apenas para registar os sucessos…
A SPE vale o que, como um todo, valerem os seus sócios. Decerto, a Sociedade Portuguesa de Estatística, onde cada sócio é parte interveniente e fundamental (pelo menos) em algum momento, continuará cada vez com mais sucesso, na certeza de que este é uma estatística "somatório" de pequenos acasos. Esteja cada um atento ao seu momento de ação!

Fernando Rosa

O Tema Central do próximo Boletim SPE será:
40 anos SPE: De onde viemos? Onde estamos? Para onde vamos?

Mensagem da Presidente

Caros sócios da SPE

Como de costume, para escrever este texto começo por ler os dois textos anteriores (lag 1 e lag 12, em terminologia de séries temporais) e sou obrigada a refletir no quão diferente é a nossa situação neste ano de 2020. Os planos para as comemorações do 40º aniversário da SPE tiveram de ser revistos em virtude da incerteza decorrente da pandemia. O XXV Congresso foi adiado e decorrerá nos dias 13 a 16 de outubro de 2021, no Hotel Évora. Agradeço desde já à Comissão Organizadora todo o trabalho desenvolvido até ao momento, com muita energia e entusiasmo e terem aceitado adiar os trabalhos de organização para o próximo ano. Agradeço ainda aos colegas da Comissão Científica que se disponibilizaram para permanecer nesta Comissão e aos oradores convidados que, entendendo a situação, se disponibilizam também para adiar a participação. Também a série "12 meses 12 iniciativas" foi suspensa como resultado da suspensão generalizada das atividades presenciais nas Instituições, mas será retomada já dia 8 de julho - notícia detalhada no Boletim.

O vírus SARS-CoV-2, a doença resultante, COVID-19 e a subsequente pandemia fizeram emergir toda uma classe de novos analistas de dados que encontraram nos media tradicionais e nas redes sociais eco fácil para as suas muito duvidosas análises. A panóplia de artigos e o grau de disparate em alguns deles levaram à escrita e publicação (ou não) de artigos de resposta com intuito pedagógico. Estes textos e outra informação relacionada com a pandemia está disponível numa entrada específica da página da SPE.

A celebração dos 40 anos da SPE não será esquecida e terá lugar a 28 de novembro em formato e com atividades a divulgar oportunamente.

Em 2020 mantém-se a atribuição dos Prémios Estatístico Júnior 2019/2020 patrocinados pelo CMUC (Centro de Matemática da Universidade de Coimbra) e o Prémio SPE 2020 que está aberto a candidaturas até 15 de setembro.

A SPE tem finalmente uma nova página web. Temos a consciência de que ainda contém gralhas que agradeço que sejam reportadas. A página tem uma área de sócios que estará ativa em breve depois de passar os testes de segurança informática. Será enviado um mail aos sócios logo que a possibilidade de registo esteja disponível.

Ainda a propósito da COVID-19, a FENStatS organizou um «COVID-19 Working Group» com o objetivo de coordenar a nível Europeu os esforços para mais e melhores dados que permitam um combate efetivo à pandemia e suas consequências quer a nível da saúde, quer da economia. A Direção nomeou a colega Laetitia Teixeira como representante da SPE no «COVID-19 WG». Agradeço desde já à Laetitia a sua disponibilidade.

Apelo aos sócios que contribuam para a série "12 meses 12 iniciativas" com ideias e iniciativas que podem ser (síncronas ou não) em forma de *webinar*. Este novo formato tem as suas vantagens pois permite a participação sem os problemas associados a deslocações. Claro que tem várias desvantagens, mas concentremo-nos nas vantagens!

Para terminar lembro que o mandato da Direção termina em dezembro pelo que durante este ano decorrerão as eleições dos Corpos Sociais para o triénio 2021-2023.

A Sociedade é dos sócios e para os sócios e é, essencialmente, o que os sócios fizerem dela. E por isso contamos com todos vós, os nossos sócios!

Até breve,

Porto, 10 de Junho de 2020

Cordiais saudações

Maria Eduarda Silva

Notícias

• Nova Programação de Atividades SPE para 2020 - 2021

A pandemia associada ao vírus SARS-Cov-2 teve consequências nas atividades da SPE programadas para este ano civil, em particular no que concerne às celebrações do 40º aniversário.

Assim a Direção vem informar sobre:

- **o adiamento do XXV Congresso da SPE -** o XXV Congresso terá lugar entre 13 e 16 de outubro 2021, Évora Hotel, Évora;
- a celebração dos 40 anos da SPE, que estava prevista decorrer durante o congresso deste ano, será comemorada num formato a divulgar posteriormente e terá lugar a 28 de novembro;
- a série 12 meses 12 iniciativas que tendo estado suspensa, será retomada com a realização do VIII Workshop of Probability and Statistics Group - Interdisciplinarity and Applications, com programa e acesso disponíveis em: https://sites.google.com/view/workshopsps-cidma, dia 8 de julho; espera-se que outras iniciativas de âmbito nacional que estavam a ser preparadas com este propósito possam ainda ser realizadas;
- a informação referente a iniciativas nacionais e internacionais sobre a pandemia COVID-19 na página web, https://www.spestatistica.pt/noticias/noticia/covid-19-e-estatistica;
- o Prémio SPE 2020 cujas candidaturas estão abertas até 15 de setembro 2020, regulamento disponível neste Boletim e na página web da SPE.

Os trabalhos com vista à criação de uma nova página não foram interrompidos. Neste momento já se encontra *on line* e agradecemos sugestões dos sócios para a sua melhoria.

A Direção da SPE

• Nova IMS e COTEC lançam modelo preditivo da evolução do COVID-19 em Portugal

A NOVA Information Management School (NOVA IMS) e a COTEC lançaram recentemente um conjunto de dashboards, capazes de prever a evolução da pandemia em Portugal. O COVID Insights (https://insights.cotec.pt/) nasce da necessidade emergente de providenciar uma visão transversal e continuamente atualizada sobre a evolução da pandemia e do seu impacto sobre os vários domínios da sociedade, desde logo a vertente epidemiológica mas também as perspetivas económica, social e demográfica, de extrema importância na retoma desejada.

A solução assenta numa plataforma analítica avançada, que integra inúmeras fontes de dados e Big Data, num Data Lake centralizado, que se diferencia de outras por ter na sua génese modelos Bayesianos hierárquicos para análise preditiva da incidência da infeção, número de recuperados e óbitos, métricas dinâmicas e ponderadas por indicadores como a densidade populacional, rendimento per capita ou mobilidade relativa e por disponibilizar visualizações avançadas, em Microsoft Power BI, que permitem correlações entre várias perspetivas e representações georeferenciadas das mesmas.

O sucesso desta solução depende dos insights obtidos e por isso a plataforma continuará a ser alimentada com novas fontes de informação, novos domínios e novas análises, todas públicas, esperando contribuir para uma tomada de decisão melhor fundamentada, tão importante no atual ambiente de incerteza.

Jorge M. Mendes

Webinar da secção de Biometria da Sociedade Portuguesa de Estatística

"Onde anda a Bioestatística na Covid-19?"

26 de junho de 2020
10h30

Moderador: Inês Sousa (Universidade do Minho)

Oradores Convidados:

"Incerteza em tempo de pandemia por SARS-CoV-2: o papel dos métodos estatísticos na vigilância epidemiológica"

Baltazar Nunes
Instituto Nacional de Saúde Doutor Ricardo Jorge
Escola Nacional de Saúde Pública da
Universidade NOVA de Lisboa

"A bioestatística como ferramenta para revelar o que há sob a ponta do iceberg - covid-19."

Milton Severo
Instituto de Saúde Pública da Universidade do Porto
Faculdade de Medicina da Universidade do Porto

Participação à distância em tempo real, através da aplicação Zoom.

Necessidade de inscrição prévia através do formulário (envio de link por email).

https://docs.google.com/forms/d/e/1FAIpQLSeR7GXONB_9usjGYBgoUFkbNIKtadJepbPIkjxyxQC1RfM2yg/viewform?usp=sf_link

Inês Sousa

• Sobre Covid-19 e a Estatística

A *Sociedade Portuguesa de Estatística* congratula-se com o empenho da comunidade científica na análise dos dados relativos à pandemia do CoVid-19, que assim contribui para um maior conhecimento da forma de propagação da doença, propiciando condições para uma atuação informada das entidades responsáveis. Diversas têm sido as intervenções feitas por qualificados estatísticos português nos diferentes órgãos de comunicação. A página da SPE *https://www.spestatistica.pt* dedica atenção especial à temática.

O *International Statistical Institute* na sua página *https://www.isi-web.org/index.php/covid-19* através do seu Comité *Vox Populi* da Estatística, apresenta alguns recursos de interesse estatístico sobre a pandemia COVID-19.
Aquela página contém informação e dá acesso a conteúdos sobre informação geral, fontes de dados, palestras e artigos, modelos e estudos de interesse geral

FR

Sobre o "*Boletim SPE*" como "*e-Boletim SPE*" – de novo o desafio

A época que vivemos, de "novo normal que não é normal", também já afetou este Boletim. A jeito de Notícia de última hora, registe-se que foi decidido que, para esta edição, apenas existisse a versão digital. Toda a logística atual – comunicação, distribuição e entrega – assim aconselha? É uma verdade que se deve acrescentar ao conjunto de debates e reflexões que em todas as sociedades, instituições e público em geral se tem feito sobre o assunto "papel versus digital". Em algumas das nossas edições também este assunto foi abordado; em particular nos editoriais de primavera 14 ou outono 16. Este Boletim primavera SPE 20, torna-se assim num *e-Boletim.* A continuação da reflexão ajudará para tomar uma decisão sobre o próximo.
Tão rápido quanto os novos tempos exigem, esta decisão difícil, pela opção feita e que ultrapassa o editor, deve aumentar o estímulo e a determinação em continuar com esta publicação histórica da *Sociedade Portuguesa de Estatística* iniciada em 1979 e que o memorial da Estatística exige e sempre espera.
Aumentar a partilha na divulgação é a metodologia que nos deve guiar para melhor cumprir os objetivo estatutários que sempre nortearam a atividade da SPE: promover, cultivar e desenvolver em Portugal, o estudo da Estatística, suas aplicações e Ciências afins (art 1, nº2).
Para esta nota, publicada como um *a posteriori* sobre o restante texto e já quando a maqueta editorial produzida na gráfica se concretizava, surgiu este "pequeno espaço" para explicar, justificar e introduzir "a versão digital".
Esta edição, *Boletim primavera 20 f*ica, assim, confirmada como *outlier a posteriori*; aquela que *a priori*, o tinha sido no Editorial.

FR

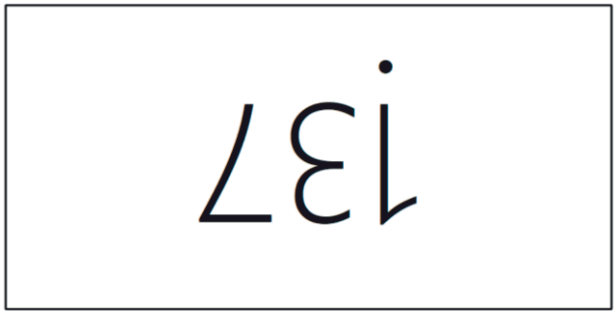
Enigmística de mefqa

AMOSTRA

$$\frac{\bar{x}}{0\quad 0} \qquad \frac{\bar{x}\quad \bar{x}}{0\qquad\quad 0} \qquad \frac{\bar{x}}{0\quad 0}$$

No Boletim SPE outono de 2019 (p. 15):

convergẽncia



convergência quase certa

lei dos grandes números

INE - 85 anos de estatísticas a servir o país

Mensagem do Presidente do Instituto Nacional de Estatística

O Instituto Nacional de Estatística, I.P. (INE) agradece à Sociedade Portuguesa de Estatística o convite para participar nesta edição do seu Boletim, em particular por esta ocasião tão especial como o 85.º aniversário do INE, comemorado no dia 23 de maio de 2020.

O INE, na sua designação atual, nasceu com a Lei n.º 1911, publicada a 23 de maio de 1935, no então denominado Diário do Governo. São 85 anos já decorridos de serviço à sociedade, procurando dar resposta às necessidades estatísticas nacionais e internacionais.
Hoje, o INE assume, entre outros, os desafios da ciência dos dados, da segurança e da tecnologia, da inovação e do manuseamento de grandes volumes de dados, do acesso a novas fontes, com o propósito de devolver à sociedade estatísticas de valor para um melhor conhecimento, investigação e a tomada de decisão.

A informação gerida pelo INE, os seus processos de suporte, sistemas, aplicações e redes são ativos valiosos para a sociedade. A garantia de confidencialidade, integridade e disponibilidade da informação assegura a credibilidade dos serviços prestados pelo INE. Neste sentido, o INE assumiu como objetivo a sistematização do seu Sistema de Gestão de Segurança de Informação e o seu alinhamento com as melhores práticas internacionais, sendo as Normas ISO referenciais já seguidos em outros domínios do INE, no âmbito da sua atividade de garantia da qualidade, aos níveis nacional e europeu.

A evolução recente da COVID-19 tem constituído para o INE, e para a comunidade estatística nacional e internacional, um enorme desafio. Além da execução do programa corrente de divulgação de resultados, já de si exigente, foram desenvolvidos novos inquéritos, novos módulos em inquéritos já existentes, exploradas novas fontes de dados, novas metodologias e novas perspetivas de análise, procurando caracterizar os impactos económicos e sociais desta pandemia. Assim, temos conseguido manter a cadência da atividade, num cenário de dificuldades de recolha de informação primária junto dos cidadãos e empresas e, simultaneamente ter a iniciativa de produzir nova informação estatística relevante para ajudar o país a enfrentar esta difícil situação.

No meio do ritmo intenso e contínuo de produção estatística do INE, também são necessários momentos em que temos de refletir continuamente sobre o que e como fazemos as estatísticas que produzimos e qual a melhor forma de as comunicar. Nem sempre é fácil num mundo onde as exigências de informação nunca abrandam e onde é necessário tomar continuamente as opções mais adequadas ao longo de toda a cadeia de produção estatística. Qual a melhor forma de recolher os dados? Qual a metodologia mais apropriada? Que métodos selecionar? Como apresentar os resultados? Como os analisar? São algumas das questões com que nos debatemos diariamente, cujas respostas não derivam de um processo automático (apesar da automatização de processos ser cada vez mais prevalente também no INE) e para as quais contribuem os mais de 600 técnicos do INE.

Estes são também tempos para repensar processos e produtos estatísticos, a relação com os fornecedores de informação (cidadãos, empresas, instituições), utilizadores (entre os quais os investigadores) e as parcerias que pretendemos ainda mais eficazes, conforme o INE preconiza nos Valores que assume.

A participação do INE neste Boletim da SPE por ocasião do nosso aniversário inclui: um texto sobre projetos de inovação em desenvolvimento; a apresentação de três trabalhos técnicos da nossa instituição; um contributo do Banco de Portugal, autoridade estatística no âmbito do Sistema Estatístico Nacional

(que acedeu ao nosso convite para colaborar nesta edição de comemoração do nosso aniversário); um contributo abordando um conjunto de projetos na área da literacia estatística em que estamos envolvidos; e por fim, a apresentação de um artigo sobre a história do nosso património arquitetónico e espólio documental.

No exercício da sua independência técnica e como autoridade estatística, a comunicação do INE procura difundir todas estas dimensões no âmbito de uma política de transparência e de proximidade tão necessária para a confiança nas estatísticas oficiais e no Sistema Estatístico Nacional.

Prof. Francisco Lima

Presidente do Conselho Diretivo do INE

Inovação nas Estatísticas Oficiais

Principais desafios do INE

A atividade do Instituto Nacional de Estatística, I.P. (INE) está enquadrada pelas estratégias dos Sistemas Estatísticos Nacional e Europeu, que dão especial relevo às inovações tecnológicas e metodológicas no processo de produção estatística nas suas diferentes fases, assim como privilegiam o acesso a novas fontes de dados, nomeadamente as de cariz administrativo. O acesso a diferentes fontes de dados e a sua integração é fundamental para conseguir desenvolver estatísticas mais relevantes, atempadas e com uma maior granularidade. No processo, consegue-se igualmente a minimização do peso da recolha junto de empresas e cidadãos, com impacto significativo na diminuição da carga estatística. A criação de um estado permanente de inovação no INE, que seja compatível com a manutenção da atividade regular, é um meio alinhado com o objetivo de beneficiar a Sociedade com informação estatística de valor acrescido, adequada às necessidades dos utilizadores para a leitura da realidade e para a tomada de decisão.
O INE tem por Missão produzir, de forma independente e imparcial, informação estatística oficial de qualidade, relevante para a Sociedade, promovendo a coordenação, a análise, a inovação e a divulgação da atividade estatística nacional, garantindo o armazenamento integrado de dados.
Neste artigo, iremos abordar três linhas de desenvolvimento associadas à área da Inovação no INE: Infraestrutura Nacional de Dados, Sistema de Gestão de Segurança de Informação e *StatsLab* (com alusão a alguns projetos associados).

Infraestrutura Nacional de Dados

A inovação tecnológica e a integração de dados de múltiplas fontes para fins estatísticos constituem os grandes desafios do INE, visando prosseguir a estratégia que tem vindo a ser concretizada nos últimos anos: um caminho no qual a informação digital passa a ter cada vez mais relevo no processo de produção estatística, seja por via da apropriação crescente de dados administrativos ou de outras fontes, seja pela adoção de processos de recolha de dados tecnologicamente mais eficientes e automatizados.
A integração de dados é o fator-chave na estratégia do INE, materializada através do desenvolvimento da Infraestrutura Nacional de Dados (IND) iniciado em 2019.
Recorrendo às competências, atribuições e missão do Instituto, a IND tem como objetivo adotar o uso mais intensivo e integrado dos dados na produção de informação estatística, aproveitando toda a cadeia produtiva das estatísticas oficiais portuguesas, desde o desenvolvimento de plataformas, aplicações e algoritmos, recolha e validação de dados, até à análise da informação estatística.
Com a intensificação da apropriação e utilização de dados administrativos e de outras fontes no processo produtivo, antecipa-se um grande aumento do volume de dados e um alargamento substancial dos domínios cobertos.
O desenvolvimento da Infraestrutura Nacional de Dados no INE tem como principais objetivos:

- Ser garante da segurança e qualidade de dados, fornecendo serviços integrados de dados, metadados e metainformação.
- Criar um único ponto de acesso aos dados administrativos e disponibilizar um conjunto de dados e recursos relacionados de modo a servir múltiplos propósitos ou projetos, independentemente de onde os dados são mantidos ou como os dados podem ser acedidos (abertos, protegidos ou seguros).
- Adotar, ao longo das cadeias de acesso, exploração e processamento de dados administrativos, mecanismos de verificação e auditoria internos e externos, eventualmente

com consagração legal, que assegurem a confiança da sociedade na gestão da infraestrutura e previnam o seu uso indevido.
- Influenciar produtores de dados públicos e privados em relação às estratégias de acesso e gestão dos dados, metodologias e tecnologias utilizadas e promover a introdução de objetivos estatísticos na produção legislativa e na regulação.
- Ser flexível para lidar com as necessidades em mudança dos seus principais utilizadores e fornecedores de dados.
- Contribuir para o desenvolvimento das estratégias de governação dos dados para fins estatísticos em Portugal.
- Promover a partilha e cooperação na recolha, acesso, transformação, processamento, validação e análise de dados.
- Diminuir a carga administrativa e estatística sobre as empresas e os cidadãos.
- Contribuir para a criação de uma Base Nacional Oficial de Moradas e para um Sistema de Informação Cadastral Simplificada.
- Contribuir para a melhoria da tomada de decisão pública, suportada em melhor informação estatística e acrescida capacidade analítica.
- Aumentar o impacto económico e social do bem público informação estatística.

Sistema de Gestão de Segurança de Informação

O processo de produção e análise estatística, a par com a utilização de quantidades consideráveis de dados administrativos nesse processo, implica a existência de uma estrutura sólida que assegura a proteção e integridade dos dados. A informação gerida pelo INE, os seus processos de suporte, sistemas, aplicações e redes são ativos valiosos para a sociedade. A garantia de confidencialidade, integridade e/ou disponibilidade da informação assegura a credibilidade da atividade do INE.
Neste sentido, o INE assumiu como objetivo a sistematização do seu Sistema de Gestão de Segurança de Informação (SGSI) e o seu alinhamento com as melhores práticas internacionais, nomeadamente cumprindo a Norma Portuguesa ISO/IEC 27001:2013, através da qual o INE se encontra certificado, no âmbito da no âmbito da proteção da informação de suporte ao processo de Micro-Data Exchange Intra-EU, do INE para o Sistema Estatístico Europeu. O SGSI é composto por um conjunto de políticas e procedimentos detalhados abrangendo todos os processos do INE, conduzindo à operacionalização do Sistema destacando-se o controlo de acessos; a classificação de confidencialidade da informação *Backup*; a transferência da informação; os controlos criptográficos; a segurança de comunicações.
Integrado no SGSI, o INE assume vários compromissos neste domínio com todos os fornecedores e utilizadores da informação. Esses compromissos, de relevância estratégica para o INE, são públicos encontrando-se disponíveis no Portal do INE, sendo materializados através de:

- A **Carta da Qualidade**, edição de 2019, que formaliza o compromisso público que o INE assume em relação à qualidade e credibilidade das estatísticas oficiais que produz e difunde, ao serviço público que presta a toda a sociedade, explicitando-o em relação aos prestadores de informação, aos utilizadores de informação estatística e a todos os cidadãos interessados e à segurança da informação;
- A **Política de Segurança da Informação**, que estabelece os princípios gerais que devem ser aplicados pelo INE, aos ativos por si geridos no âmbito do SGSI, alinhada com os requisitos da NP ISO/IEC 27001:2013, a legislação e regulamentação aplicáveis e as recomendações do Sistema Estatístico Europeu e do EUROSTAT, específicas em matéria de segurança da informação;
- A **Política de Confidencialidade Estatística**, que constitui o compromisso público quanto à observância do Princípio do Segredo Estatístico, assumido pelo INE, enquanto órgão central responsável pela coordenação e desenvolvimento da atividade estatística nacional;
- A **Política de Privacidade e Proteção de Dados Pessoais**, que visa fornecer ao titular dos dados informações sobre a natureza dos dados recolhidos, a respetiva finalidade e o tratamento que será realizado.

***StatsLab* (estatísticas em desenvolvimento)**

No âmbito da inovação, destaca-se o *StatsLab*, um espaço dedicado à apresentação de estatísticas em desenvolvimento disponível no Portal do INE. Estas estatísticas distinguem-se por duas características: (i) inserem-se em projetos de novos produtos estatísticos que ainda não foram inteiramente completados e, contudo, (ii) expressam já informação que se pode revelar útil para a análise económica e social.
A possibilidade crescente de acesso pelo INE a fontes administrativas e a fontes não convencionais, designadamente obtidas junto de entidades privadas, colocam novos desafios à produção das estatísticas pelo INE. O *StatsLab* – Estatísticas em Desenvolvimento é um espaço onde são apresentados novos produtos estatísticos antes de adquirirem o seu formato final e visando tirar partido dessas fontes.
Adicionalmente serão também testadas novas formas de apresentação de informação e resultados derivados de estatísticas já atualmente publicadas, explorando novas ferramentas analíticas.
Atualmente o *StatsLab* integra os seguintes conteúdos:

- **Censos com dados Administrativos -** divulgação do progresso da linha de investigação Censos com base em dados administrativos, após os Censos 2021. Este projeto insere-se no quadro de desenvolvimento da Infraestrutura Nacional de Dados que dá corpo à estratégia do INE de integração e criação de valor para a sociedade a partir de diferentes fontes de dados. Central ao projeto é a constituição da Base de População Residente que cobre um conjunto de características – geográficas, demográficas e socioeconómicas – da população residente em Portugal e que resulta da integração de informação administrativa proveniente de diversas fontes da administração pública.
  A Base de População Residente é construída através da aplicação de técnicas de *record linkage* e *matching* tendo em vista a integração de informação administrativa de diferentes fontes administrativas. Portugal não dispõe de um número de identificação único, utilizado transversalmente pelas várias entidades da Administração Pública, colocando desafios acrescidos à integração da informação das diferentes bases de dados administrativas. Em particular, é necessário determinar se uma pessoa reside no território nacional, o que corresponde ao conceito de população residente associado às operações censitárias. Para chegar a este conceito é aplicado um conjunto de regras designadas de "indícios de residência". Estas regras permitem validar a residência em Portugal através da presença do indivíduo nas diferentes bases de dados administrativas (e.g., o indivíduo trabalha, frequenta o sistema de ensino, paga impostos, está inscrito no centro de emprego).

- **Inquérito ao Setor da Economia Social 2018 -** apresentação dos principais resultados do Inquérito ao Setor da Economia Social, por ocasião do Dia Europeu das Empresas da Economia Social. Trata-se de um inquérito realizado pela primeira vez no âmbito do Sistema Estatístico Nacional, promovido pelo INE em colaboração com a CASES para apurar informação sobre caracterização deste setor em 2018. Os resultados apresentados centram-se essencialmente na análise das práticas de gestão das entidades da Economia Social, agrupadas em 5 grandes famílias – Cooperativas, Associações Mutualistas, Misericórdias, Fundações e Associações com fins altruísticos. Pretende-se num futuro próximo um desenvolvimento dos resultados, nomeadamente através da divulgação de informação que permita uma caracterização mais detalhada do setor, em termos das atividades desenvolvidas, composição interna, relações com entidades do setor público e privado, indicadores de medição do impacto social destas entidades e modalidades de financiamento.

- **Estatísticas do Rendimento ao nível local –** indicadores de rendimento declarado no IRS: divulgação das "Estatísticas do Rendimento ao nível local" com base em dados fiscais anonimizados da Autoridade Tributária e Aduaneira relativos à Nota de liquidação do Imposto sobre o Rendimento das Pessoas Singulares (IRS – Modelo 3), obtidos no âmbito de um protocolo celebrado entre a Autoridade Tributária e o INE.

- **Remuneração bruta mensal média por trabalhador –** cálculos do INE com base na informação da Segurança Social e da Caixa Geral de Aposentações: divulgação

trimestralmente, estatísticas sobre remunerações, com base na informação da Declaração Mensal de Remunerações transmitidas pelas empresas à Segurança Social e da Relação Contributiva dos subscritores da Caixa Geral de Aposentações. A informação cobre cerca de 400 mil empresas e a aproximadamente 4,2 milhões de trabalhadores. Pretende-se, no futuro, quando o INE dispuser de informação ao nível do trabalhador, criar estatísticas que permitam, entre outras possibilidades, conhecer a distribuição das remunerações e proceder à caracterização sociodemográfica dos trabalhadores.

- **Indicadores de mobilidade da população ao nível regional -** uma leitura a partir da informação da iniciativa "*Data for Good*" do Facebook sendo divulgados indicadores de mobilidade da população ao nível das NUTS III no território nacional. Estes dados correspondem a atualizações de localização recolhidas a partir dos dispositivos móveis de utilizadores da aplicação Facebook que têm a opção "histórico de localização" ligada.

Referências:

- Plano de Atividades do INE e das Entidades com Delegação de Competências 2020, INE, 2020;
- Carta da Qualidade, 5.ª edição, INE, 2019;
- Política de Segurança da Informação, INE, 2019;
- Política de Confidencialidade Estatística, INE, 2019;
- Política de Privacidade e Proteção de Dados Pessoais, INE, 2019.

Modelos de regressão para dados de contagem: uma aplicação ao Inquérito ao Transporte Rodoviário de Mercadorias

Inês Rodrigues, *ines.rodrigues@ine.pt*

Departamento de Metodologia e Sistemas de Informação, Instituto Nacional de Estatística

Introdução

O *web scraping* – técnica que permite a extração e armazenamento automáticos de dados a partir de páginas *web* [1] – tem sido regularmente utilizado pelo INE no âmbito do Inquérito ao Transporte Rodoviário de Mercadorias (ITRM) [2]. A obtenção de dados fiáveis e atualizados sobre a situação das matrículas permite atualizar a base de amostragem e amostra em relação às matrículas canceladas, com a consequente melhoria nas taxas de resposta. Por outro lado, promove um desafio: na ausência de informação igualmente atualizada sobre os registos de novas matrículas, a exclusão das matrículas canceladas da base de amostragem, ao longo do ano, conduz necessariamente à diminuição da sua dimensão, com efeitos sobre o cálculo dos ponderadores e estimativas trimestrais. De modo a estudar a possibilidade de implementação de um fator de correção para a entrada de matrículas na base de amostragem, procedeu-se ao estudo de modelos que permitam estimar o número anual de novas matrículas, por estrato.

O Inquérito ao Transporte Rodoviário de Mercadorias

O Inquérito ao Transporte Rodoviário de Mercadorias (ITRM) visa a produção de informação sobre o tráfego de mercadorias por estrada e suas principais características: capacidade e grau de utilização do parque nacional de veículos matriculados em Portugal continental, fluxos de tráfego e natureza das mercadorias. O ITRM é um inquérito amostral de periodicidade trimestral, com unidade estatística de observação correspondente ao "veículo pesado de mercadorias": camiões e tratores rodoviários em atividade, com peso bruto (camiões) ou tara (tratores rodoviários) superior a 3 500 Kg. Com vista ao dimensionamento e seleção da amostra, a base de amostragem do ITRM é estratificada pelo cruzamento das variáveis região do Continente (NUTS II), categoria do veículo (camião/trator), escalão de peso bruto/tara e tipo de parque (conta própria/conta de outrem), num total de 70 estratos [2].

Identificação de matrículas canceladas via *web scraping* no website do IMT, I.P.

O *web scraping* é uma técnica computacional que automatiza o processo de recolha de dados a partir de páginas web, através de ferramentas que navegam e extraem a informação semiestruturada e a armazenam em bases de dados [1]. Em junho de 2015, iniciou-se a aplicação de *web scraping* ao *website* do Instituto da Mobilidade e dos Transportes (IMT, I.P.), com vista à atualização mais frequente e acessível da informação relativa à situação do veículo; em particular, esta técnica tem sido utilizada regularmente, com frequência trimestral, para identificação das matrículas canceladas. Em cada momento, é verificada a situação das matrículas ativas; no final de cada ano é adicionalmente verificado o estado das matrículas reprovadas na inspeção. A informação obtida é utilizada no momento de seleção da amostra de um dado trimestre (não sendo selecionadas matrículas canceladas), bem como para conseguir informação sobre as matrículas canceladas que se encontrem em amostras de trimestres anteriores.

Correção da base de amostragem para considerar a entrada de novas matrículas

A informação relativa ao registo de novas matrículas é enviada anualmente ao INE pelo Instituto dos Registos e do Notariado, I.P., sendo considerada no momento de constituição do universo de veículos. No decorrer do período de recolha, não existem atualizações sobre as matrículas que foram entretanto registadas. Assim, dado que a dimensão inicial da base de amostragem é corrigida para considerar a saída dos veículos identificados ao longo da recolha como fora de âmbito, o facto de não se considerar a possibilidade de entrada de novas matrículas pode resultar na subestimação dos valores apurados, bem como na diminuição dos valores estimados em diferentes momentos ao longo do tempo. De modo a estudar a possibilidade de implementação de um fator de correção para a entrada de novas matrículas na base de amostragem, procedeu-se ao estudo de modelos que permitam estimar o número anual de novas matrículas, por estrato.

A estimação do número de novas matrículas em cada ano foi baseada na comparação entre as bases de amostragem de dois anos consecutivos, entre o final de 2011 e o final de 2017. Assume-se que as matrículas presentes na base de amostragem do ano $x + 1$ e que não estavam incluídas na base de amostragem do ano x correspondem a novos registos. No conjunto dos anos considerados (2012 a 2017, como anos completos) foram registadas 46 932 novas matrículas.

Os resultados apresentados no Quadro 1 sugerem a existência de diferenças na entrada de novos registos, entre as diferentes categorias das variáveis de estratificação. Tendem a ser registados mais veículos novos na região da Área Metropolitana de Lisboa e menos no Algarve; entre os camiões, são registados mais veículos de peso bruto entre 3 501 e 10 000 Kg, enquanto entre os tratores são registados mais veículos com tara superior a 7 000 Kg; tendem ainda a ser registados mais veículos do parque por conta própria do que por conta de outrem.

Quadro 1 - Matrículas novas, por estrato, por variável de estratificação, 2012-2017

	Mín	Q1	Mediana	Média	Q3	Máx	Var	Soma
Região NUTS II								
Norte	5	22,3	63,5	153,0	234,5	914	36 332,0	12 852
Centro	5	22,8	57,5	169,2	174,3	1 574	67 435,5	14 210
A.M.Lisboa	2	21,8	62,0	187,7	229,8	1 627	79 509,9	15 767
Alentejo	0	8,0	19,0	37,7	45,3	191	2 024,0	3 169
Algarve	0	2,0	5,0	11,1	13,3	70	236,0	934
Tipo de veículo e escalão de peso bruto/tara								
Camião								
3 501 - 10 000 Kg	0	20,0	79,0	172,4	277,3	714	40 120,5	10 344
10 001 - 16 000 Kg	0	7,8	20,5	32,3	46,5	175	1 241,0	1 939
16 001 - 19 000 Kg	1	9,8	26,5	40,4	59,5	177	1 521,3	2 424
19 001 - 26 000 Kg	0	7,5	19,0	32,4	45,3	163	1 304,6	1 943
Mais de 26 000 Kg	0	4,8	11,0	13,3	16,3	54	142,4	798
Trator								
3 501 - 7 000 Kg	2	57,0	175,5	219,3	348,0	809	39 158,6	13 157
Mais de 7 000 Kg	2	19,8	99,0	272,1	420,5	1 627	14 5560,5	16 327
Tipo de Parque								
Por conta própria	0	14,0	39,0	121,7	121,3	1 627	41 539,4	25 559
Por conta de outrem	0	8,0	20,5	101,8	86,5	1 574	42 543,1	21 373

Com vista a analisar com maior detalhe as diferenças entre estratos no que diz respeito à entrada de novas matrículas e a estudar a possibilidade de estimar a entrada de novos registos para anos futuros, procedeu-se à estimação de modelos de regressão, com variável dependente correspondente ao número anual de novas matrículas, por estrato, entre 2012 e 2017.

Definindo a entrada de uma nova matrícula no universo como o evento de interesse e considerando que a variável dependente se refere à contagem de eventos ocorridos, por estrato, ao longo de cada unidade

de tempo (um ano), a tentativa inicial de modelar esta variável baseia-se num modelo de regressão de Poisson [3]. Sendo as contagens relativas ao estrato h ($h = 1, \dots, H$) representadas por Y_h, a dependência do valor médio condicional de Y_h, $E[Y_h|X_h] = \mu_h$, sobre as covariáveis X_h, é especificada através de uma relação log-linear, como:

$$\log(\mu_h) = \eta_h = \mathrm{X}_h^T\beta$$

em que β representa o vetor de coeficientes de regressão associados a X_h. Importa destacar que a modelação das contagens com base na distribuição de Poisson implica assumir, contudo, que $Var(Y_h|X_h) = \mu_h$. Através da análise descritiva já realizada, é possível supor que esta hipótese seja demasiado restritiva no caso presente e que a relação $Var(Y_h|X_h) > \mu_h$ parece ser mais adequada.

Uma das formas de modelar a sobredispersão dos dados passa por manter a função de regressão do modelo de Poisson para a média, mas substituir a relação média-variância por $Var(Y_h|X_h) = \phi\mu_h$. Esta opção não resulta num modelo de probabilidade específico e completo para a variável resposta e, como tal, o modelo não apresenta uma função de verosimilhança totalmente especificada. É, por isso, habitualmente designado por modelo de quasi-Poisson [4].

Uma outra alternativa consiste em assumir que as contagens sobredispersas seguem uma distribuição binomial negativa. Esta distribuição obtém-se assumindo que $Y_h \sim Pois(K_h)$ onde K_h são variáveis aleatórias com distribuição gama, $\Gamma(\theta, \lambda_h)$. Como tal, $E[Y_h|X_h, K_h] = \theta/\lambda_h = \mu_h$ e $Var(Y_h|X_h, K_h) = \mu_h + \mu_h^2/\theta$. Neste caso, assume-se também uma relação log-linear entre μ_h e as covariáveis X_h [3,4].

O Quadro 2 apresenta os resultados de três modelos de regressão para o número anual de novas matrículas, com variáveis explicativas correspondentes às variáveis de estratificação consideradas no ITRM: (1) modelo de Poisson; (2) modelo de quasi-Poisson; e (3) modelo de regressão binomial negativo. Os modelos 1 e 2 foram estimados com recurso à função **glm** do package **stats** [5], e o modelo 3 através da função **glm.nb** do package **MASS** [6] do software R, versão 3.5.2.

Quadro 2 – Modelos de regressão de Poisson, quasi-Poisson e binomial negativo para estimação do número de novas matrículas, por estrato

	Modelo 1 Poisson β (se)	Modelo 2 Quasi-Poisson β (se)	Modelo 3 Bin.Negativo β (se)
Constante	5,37 (0,01)***	5,37 (0,12)***	5,2 (0,14)***
Região (NUTS II)			
Norte	ref.	ref.	ref.
Centro	0,1 (0,01)***	0,1 (0,11)	0,08 (0,13)
AML	0,2 (0,01)***	0,2 (0,1)*	0,07 (0,13)
Alentejo	-1,4 (0,02)***	-1,4 (0,17)***	-1,28 (0,13)***
Algarve	-2,62 (0,03)***	-2,62 (0,29)***	-2,39 (0,13)***
Tipo de veículo e escalão de peso bruto/tara			
Camião			
3 501 - 10 000 Kg	ref.	ref.	ref.
10 001 - 16 000 Kg	-1,67 (0,02)***	-1,67 (0,21)***	-1,63 (0,15)***
16 001 - 19 000 Kg	-1,45 (0,02)***	-1,45 (0,2)***	-1,35 (0,15)***
19 001 - 26 000 Kg	-1,67 (0,02)***	-1,67 (0,21)***	-1,64 (0,15)***
Mais de 26 000 Kg	-2,56 (0,04)***	-2,56 (0,32)***	-2,35 (0,16)***
Trator			
3 501 - 7 000 Kg	0,24 (0,01)***	0,24 (0,11)*	0,23 (0,15)
Mais de 7 000 Kg	0,46 (0,01)***	0,46 (0,11)***	0,45 (0,15)**
Tipo de Parque			
Por conta própria	0,18 (0,01)***	0,18 (0,08)*	0,45 (0,08)***
Por conta de outrem	ref.	ref.	ref.
nº parâmetros	12	12	13
log L	-16 020	-	-2 022
AIC	32 063,37	-	4 070,48

*** p < 0,001; ** p < 0,01; * p < 0,05

Os coeficientes estimados não diferem consideravelmente entre os três modelos. A principal distinção refere-se aos erros padrão obtidos: a possibilidade de lidar com a sobredispersão dos dados nos modelos 2 e 3 resulta em desvios bastante superiores aos estimados com base no modelo 1. A subestimação dos erros padrão no modelo 1 pode conduzir à incorreta avaliação da significância dos parâmetros – o que

poderá ocorrer no caso concreto dos coeficientes associados às regiões Centro e Área Metropolitana de Lisboa (em comparação com a região Norte).
O aumento verificado no valor do logaritmo da verosimilhança, quando comparamos o modelo 3 com o modelo 1, comprova a importância de considerar a sobredispersão dos dados. O mesmo se conclui através da comparação do AIC (*Akaike information criterion*).
O modelo 2 não tem uma função de verosimilhança associada; como tal, a decisão entre este modelo e o modelo 3 deve basear-se na principal distinção entre ambos: a relação estimada entre média e variância da variável resposta, que se assume como linear no modelo 2 e como quadrática no modelo 3. O Gráfico 1 apresenta a variância em função da média estimada, de acordo com os dois modelos. Estas funções são comparadas com os pontos correspondentes aos valores de média e variância calculados para grupos de observações de novas matrículas, constituídos com base nos percentis do valor estimado por um dos modelos. Estes resultados sugerem que o modelo de regressão binomial negativo – modelo 3 – permite captar com maior precisão a relação entre a média e a variância dos dados observados.

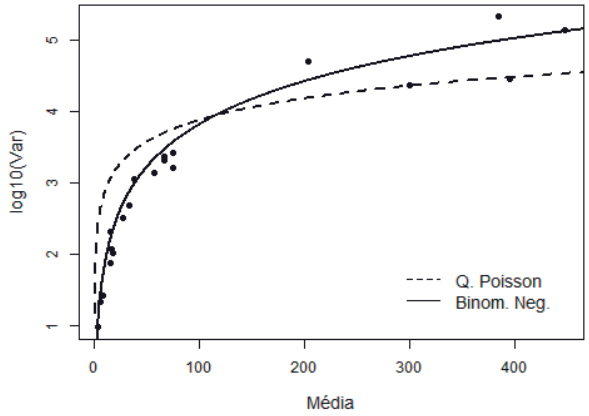


Gráfico 1 – Relação média-variância estimada, modelo de regressão quasi-Poisson e binomial negativo

Os diagramas de dispersão dos resíduos face aos valores estimados por cada modelo permitem também confirmar que o modelo 3 é o mais adequado (Gráfico 2). Contudo, os resíduos obtidos através do modelo 3 apresentam ainda um ligeiro padrão. Interessa investigar a possibilidade de modelar, adicionalmente, as interações verificadas entre as variáveis de estratificação. Foram inicialmente comparados três modelos baseados no modelo 3, ao qual foram adicionadas, separadamente, as interações possíveis entre as três covariáveis: Região x Tipo/Peso (modelo 3.1); Região x Tipo Parque (modelo 3.2); e Tipo/Peso x Tipo Parque (modelo 3.3). De acordo com os resultados obtidos, o modelo 3.3 parece representar uma melhoria relativamente ao modelo sem interações (*Likelihood Ratio Test*, LRT = 104,98; p < 0,001). De acordo com este modelo, o número esperado de matrículas novas por tipo e escalão de peso bruto/tara do veículo varia dependendo do tipo de parque ao qual o veículo pertence.

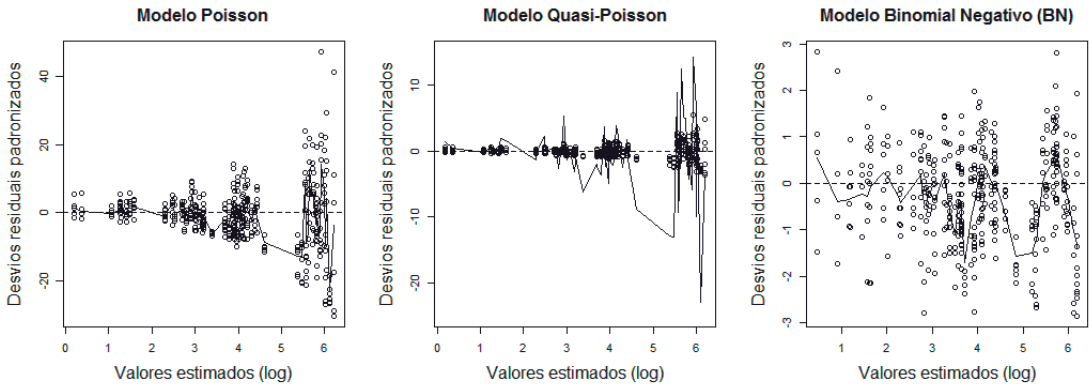


Gráfico 2 – Desvios residuais vs valores estimados, modelos de regressão de Poisson, quasi-Poisson e binomial negativo

Quadro 3 – Modelos de regressão binomiais negativos com tendência para estimação do número de novas matrículas, por estrato

	Modelo 3.3a Bin.Negativo β (se)	Modelo 3.3b Bin.Negativo β (se)	Modelo 3.3 Bin.Negativo β (se)
Constante	3,88 (0,16)***	3,77 (0,19)***	4,3 (0,15)***
Região (NUTS II)			
Norte	ref.	ref.	ref.
Centro	-0,01 (0,11)	0 (0,1)	-0,02 (0,11)
AML	0,15 (0,11)	0,13 (0,1)	0,15 (0,11)
Alentejo	-1,32 (0,11)***	-1,31 (0,11)***	-1,35 (0,11)***
Algarve	-2,44 (0,12)***	-2,42 (0,11)***	-2,42 (0,12)***
Tipo de veículo e escalão de peso bruto/tara			
Camião			
3 501 - 10 000 Kg	ref.	ref.	ref.
10 001 - 16 000 Kg	-0,92 (0,19)***	-0,69 (0,26)**	-0,96 (0,2)***
16 001 - 19 000 Kg	-0,38 (0,19)*	-0,41 (0,26)	-0,4 (0,19)*
19 001 - 26 000 Kg	-0,85 (0,19)***	-0,61 (0,26)*	-0,91 (0,2)***
Mais de 26 000 Kg	-1,25 (0,19)***	-0,92 (0,27)***	-1,33 (0,2)***
Trator			
3 501 - 7 000 Kg	1,45 (0,18)***	1,87 (0,25)***	1,39 (0,19)***
Mais de 7 000 Kg	1,88 (0,18)***	1,32 (0,25)***	1,85 (0,19)***
Tipo de Parque			
Por conta própria	1,93 (0,18)***	1,96 (0,17)***	1,85 (0,19)***
Por conta de outrem	ref.	ref.	ref.
Tipo/Peso x Tp.Parque			
Camião E2 - C. própria	-0,98 (0,26)***	-1,03 (0,25)***	-0,95 (0,27)***
Camião E3 - C. própria	-1,5 (0,26)***	-1,52 (0,25)***	-1,44 (0,27)***
Camião E4 - C. própria	-1,1 (0,26)***	-1,13 (0,25)***	-1,03 (0,27)***
Camião E5 - C. própria	-1,7 (0,27)***	-1,71 (0,26)***	-1,59 (0,28)***
Trator E1 - C. própria	-1,94 (0,25)***	-2,01 (0,24)***	-1,92 (0,26)***
Trator E2 - C. própria	-2,97 (0,25)***	-3,21 (0,24)***	-2,8 (0,26)***
Ano'	0,14 (0,02)***	0,17 (0,05)***	
Tipo/Peso x Ano'			
Camião E2 - Ano'		-0,08 (0,07)	-0,95 (0,27)***
Camião E3 - Ano'		0,01 (0,07)	-1,44 (0,27)***
Camião E4 - Ano'		-0,09 (0,07)	-1,03 (0,27)***
Camião E5 - Ano'		-0,13 (0,08)	-1,59 (0,28)***
Trator E1 - Ano'		-0,16 (0,07)*	-1,92 (0,26)***
Trator E2 - Ano'		0,24 (0,07)***	-2,8 (0,26)***
nº parâmetros	20	26	19
log L	-1 952	-1 936	-1 970
θ (se)	2,17 (0,16)	2,35 (0,18)	1,99 (0,14)
AIC	3 943,77	3 924,18	3 977,5

*** p < 0,001; ** p < 0,01; * p < 0,05

Considerando que o presente estudo tem em vista propor um modelo no qual se possa vir a basear a previsão do número de novas matrículas, interessa também estudar a possibilidade de modelar a variação temporal observada. Nesse sentido, foi adicionado ao modelo 3.3 um termo correspondente ao número de anos desde o início do período em análise (Ano' = Ano – 2012). A introdução deste termo conduziu à redução do AIC para 3 943,77 e ao aumento do logaritmo da verosimilhança para -1 952 (modelo 3.3a), revelando uma melhoria significativa na qualidade do modelo (LRT = 35,7; p < 0,001). Por fim, foi ainda avaliada a possibilidade de modelar as diferenças na evolução temporal do número de matrículas, por categoria das variáveis de estratificação. Foram adicionados, em modelos distintos, termos correspondentes à interação entre cada uma das variáveis e o ano. Apenas a inclusão da interação entre o tipo/peso do veículo e o ano conduziu a uma melhoria significativa do ajustamento,

comparativamente ao modelo sem interações com o ano (LRT = 31,6; p < 0,001). Os resultados obtidos para cada um destes modelos são apresentados no Quadro 3.
Também os diagramas de dispersão dos resíduos face aos valores estimados por cada modelo confirmam a melhoria obtida com as alterações realizadas (Gráfico 3).

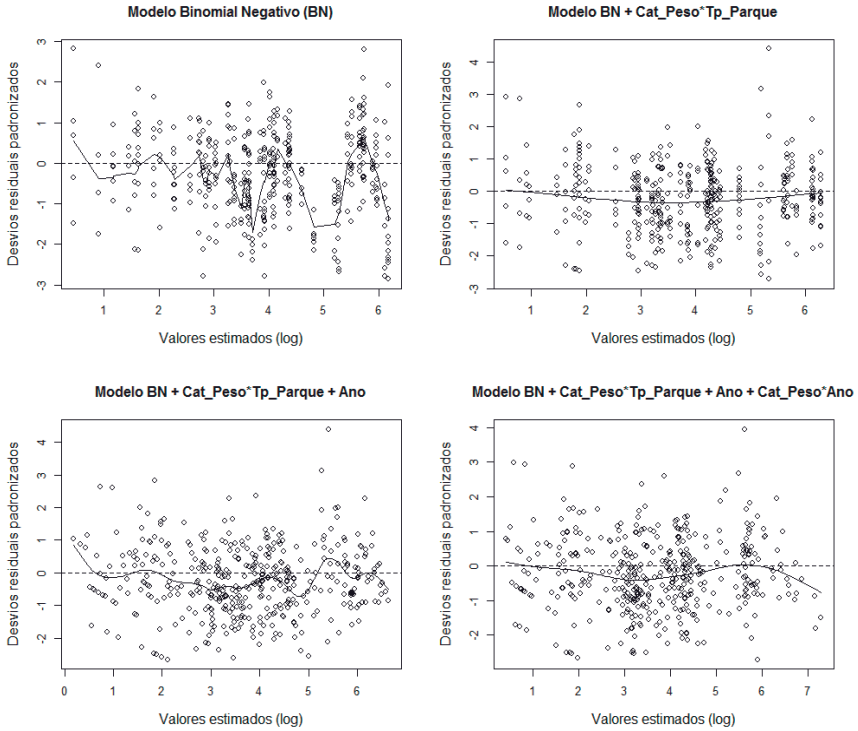


Gráfico 3 – Desvios residuais vs valores estimados, modelos de regressão binomial negativo

Notas finais

Os seguintes pontos devem ainda ser considerados nesta análise:

- Os modelos utilizados no estudo realizado até ao momento assumem que a entrada de novas matrículas em cada estrato ocorre de modo independente; contudo, é conveniente analisar até que ponto as entradas poderão estar associadas por corresponderem a veículos registados sob a mesma empresa; é de esperar que veículos correspondentes à mesma empresa tenham características semelhantes (e, como tal, pertençam ao(s) mesmo(s) estrato(s));
- Os modelos apresentados procuram estimar o número agregado de novas matrículas registadas em cada estrato; contudo, a modelação do comportamento "individual", em termos do número de veículos registados por empresa, poderá representar uma melhoria dos resultados obtidos.

Referências Bibliográficas

[1] Fernandes, M. J. (2019). O uso do web scraping nas Estatísticas Oficiais. Em: H. Bacelar-Nicolau, F. Sousa, C. Marcelo, A. S. Ferreira, P. Infante, A. Figueiredo (Eds.), *Classificação e Análise de Dados – Métodos e Aplicações III* (pp. 169-179), 1ª Edição. INE, Lisboa. Abril de 2019. URL: http://www.clad.pt/DOC_EVENTOS/CLADMAp_III_cores_online_proteg.pdf.

[2] Documento Metodológico - Inquérito ao Transporte Rodoviário de Mercadorias. Versão 4.0, DEE/CTT, INE. Janeiro de 2017. URL: http://smi.ine.pt/UploadFile/Download/2083.

[3] Rodríguez, G. (2007). Lecture Notes on Generalized Linear Models. URL: http://data.princeton.edu/wws509/notes/.

[4] Hoef, J. & Boveng, P. (2007). Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*. 88(11):2766-72. URL: https://doi.org/10.1890/07-0043.1.

[5] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

[6] Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S*, Fourth edition. Springer, New York. ISBN 0-387-95457-0, URL: http://www.stats.ox.ac.uk/pub/MASS4.

Short-Term Regional Demographic Forecasts with Time Series Methods and Machine Learning Algorithms

Jorge M. Bravo[1], *jbravo@novaims.unl.pt*
Universidade Nova de Lisboa (NOVA IMS) & MagIC & CEFAGE-EU & Université Paris-Dauphine PSL

Edviges Coelho, *edvides.coelho@ine.pt*
Departamento de Estatísticas Demográficas e Sociais, Instituto Nacional de Estatística & Universidade Lusófona (ECEO-UHLT)

1 Introduction

Forecasts of births and deaths are a critical input in the computation of resident population estimates since they determine, together with net migration, the dynamics of both the population size and its age distribution in the territory. In this study we evaluate the short-term forecasting accuracy of alternative traditional linear and non-linear seasonal time series methods (SARIMA, Holt-Winters, State Space models) and new advanced machine learning algorithms (Artificial Neural Networks, Bagging) to birth and death monthly forecasting at the sub-national level using a backtesting cross-validation approach. We use a time series of monthly data from 2000 to 2018 disaggregated by sex for the 25 Portuguese NUTS3 regions. Our results provide valuable tools for policymakers in assessing how changes in population size and age composition affect economic outcomes and their geographical patterns and in developing, implementing and coordinating active regional policy instruments.

Empirical evidence shows that there is a territorial dimension of demographic change. In recent years there is an increasing acknowledgment that the local and regional levels provide a more suitable ground for designing and implementing policy responses to the complex interaction of factors that dictate the wide-ranging patterns of demographic change (EU, 2016). Population forecasts are widely used for analytical, planning and policy purposes (e.g., education, health, housing, pensions, labour market, security, spatial planning, transportation, public infrastructure and social policy planning) at national, regional and local levels (Bravo, 2016, 2019; Bravo et al., 2018, 2020; Ayuso, Bravo & Holzmann, 2020; Bravo & Herce, 2020). Concerns about the possible long-term effects of demographic change on population size, dynamics and structure increased the importance of producing accurate population projections at the subnational level. The population of a given territorial area and its age distribution changes over time through the interaction of three possibly correlated factors: fertility, mortality, and (international and regional) migration. To project the population size and age structure at a future date, economists and demographers typically using the cohort-component method and stochastic time series methods to project the dynamics of the three components of demographic change. Forecasts of monthly births and deaths are a critical input in the computation of monthly estimates of resident population (MERP).[2]

Birth and death forecasts can in principle be produced using, among others, statistical time series methods (univariate or multivariate), structural models (e.g., VAR models) or machine learning methods (e.g., Artificial Neural Network (ANN), Support Vector Machines (SVM)). Births and deaths time series are typically non-stationary, contain a trend and exhibit strong seasonality patterns at both national and

[1] This communication brief is an abridged and updated version of a research paper presented at the ASMDA, 26th APDR and CAPSI 2019 Conferences.

[2] To produce MERP, for each subpopulation and gender it is necessary to: (i) obtain monthly forecasts of the total number of births and deaths, (ii) estimate age-specific mortality rates considering period/cohort life tables derived from stochastic mortality models, eventually considering for heterogeneity in longevity (Ayuso, Bravo & Holzmann, 2017a,b), (iii) estimate the level and age pattern of net international migration, and (iv) consider a number of assumptions such as the distribution of age-specific fertility rates or the sex ratio a birth (Bravo

regional levels. For vital events computed for small populations on monthly time intervals, the need to uncover complex structures of temporal interdependence in time series data is critically challenged in the presence of seasonal variability. In recent decades a substantial amount of research has focused on the development and application of traditional time series models in population forecasts, focusing either on total population growth or on individual components of growth (see, e.g., Pflaumer, 1992; Lee 1992; Lee and Tuljapurkar 1994; Keilman, Pham & Hetland, 2002; Booth, 2006; Tayman, Smith, and Lin 2007; Alho, Bravo and Palmer, 2012; Abel et al. 2013; Wiśniowski et al., 2015; Bravo and El Mekkaoui de Freitas, 2018; Li et al., 2018). The main focus of these studies is largely on the identification and measurement of uncertainty in population forecasts, with little interest in the assessment of the models forecasting accuracy or the out-of-sample validity of the prediction intervals. Much of the research concerning the evaluation of time series models for birth and death forecasting has been focused on univariate time series ARIMA models at the national level, with little research on the predictive accuracy of these models at the sub-national level, particularly in small population areas. Fewer still have explored the use of the Holt-Winters exponential smoothing (HW) and State Space (SS) time series models in small population exercises. Up to our knowledge, no attempt has been made to use machine learning and deep learning methods to forecasts monthly demographic data.

In this study, we address this gap and investigate and compare the predictive accuracy of alternative linear and non-linear traditional time series models (seasonal ARIMA, HW and SS) and more advanced machine learning time series methods (ANN, Bootstrapp Aggregating or Bagging) to birth and death monthly forecasting at the sub-national level using up-to-date demographic data. Using a series of monthly data from 2000 to 2018 disaggregated by sex for the 25 Portuguese NUTS3 regions, we compare the short-term (one year) method's forecasting accuracy. We adopt a backtesting time series cross-validation approach, i.e., we consider a multi-step forecasting approach with re-estimation in which the training data or base period is extended before re-selecting and re-estimating the model at each iteration and computing forecasts. Our main contributions are the following. First, we summarise and analyse the out-of-sample error performance of commonly used Seasonal ARIMA, HW and SS forecasting models together with new powerful machine learning algorithms, using a rich and large set of subpopulations and two different demographic events with different dynamics over time. Second, we evaluate the out-of-sample performance of the prediction intervals produced by these models. Third, we assess the consistency of the predictive performance of these methods in populations of different size and nature. Fourth, we evaluate the existence of significant differences in the model's forecasting accuracy between subpopulations of different sex. Fifth, we investigate how well the models perform in terms of predicting the uncertainty of future monthly birth and death counts.

The selection of the appropriate forecasting method depends on several factors, including the past behaviour pattern of the time series, previous knowledge about the nature of the phenomenon being studied, the availability of statistical data and the predictive capacity of the model. Our results show that these simulations provide valuable insights regarding the forecasting performance of alternative time series models in small population forecasting exercises and on the validity of using such models as predictors of population forecast uncertainty and, thus, have significant practical implications in territorialising public policies to address demographic change. The remaining part of the study is organised as follows. Section 2 describes the materials and methods used in this study. Section 3 details the research design used to produce forecasts and assess model's performance. Section 4 presents and discusses the results. Section 5 concludes.

2 Materials and Methods

2.1. Seasonal ARIMA Model

The seasonal ARIMA model is an extension to the classical ARIMA model that supports the direct modelling of both the trend and seasonal components of a time series and it is widely used for forecasting. The model includes new parameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality (Hyndman and Athanasopoulos, 2018). In this study, we combine the seasonal and non-seasonal components into a multiplicative seasonal autoregressive moving average model, or SARIMA model, given by

$$\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d x_t = \delta + \Theta_Q(B^s)\theta(B)w_t \quad (1)$$

where w_t denotes the Gaussian white noise process. The general model can be expressed as $ARIMA(p,d,q)\times(P,D,Q)_s$, where the ordinary autoregressive (AR) and moving average (MA) components are represented by polynomials $\phi(B)$ and $\theta(B)$ of orders p and q, respectively, the seasonal AR and MA components are denoted by $\Phi_P(B^s)$ and $\Theta_Q(B^s)$ of orders P and Q, respectively. The non-seasonal and seasonal difference components are represented by $\nabla^d = (1-B)^d$ and $\nabla^D = (1-B^s)^D$, respectively. The seasonal period s defines the number of observations that make up a seasonal cycle (e.g., $s = 12$ for monthly observations).

The estimation process for the parameters in (1) for each of the 100 time series follows the standard Box-Jenkins methodology in an iterative 3-step procedure comprising the identification, estimation and evaluation and diagnostic analysis stages, testing for unit roots and white noise errors and optimizing a stepwise algorithm for the AIC Criterion. When the data suggest the inexistence of seasonal unit roots in the series and the seasonality is deterministic, we can express it as a function of seasonal dummy variables (and time eventually). In this case, an ARIMA model if fitted to the residuals of the equation:

$$Y_t = \alpha + \sum_{i=1}^{s-1}\gamma_{i,t}D_{i,t} + \beta t + \epsilon_t \quad (2)$$

where Y_t is the variable of interest, $D_{i,t}$ are seasonal dummies, t denotes time and ϵ_t is a white-noise error term. Additionally, we examined the residuals of the selected model and formally examined the null hypothesis of independence of the residuals using the Box-Pierce/Ljung-Box test. We also tested the normality of the residuals using the Jarque-Bera Test. After examining different models, the best SARIMA model was selected, parameters were estimated using the nonlinear least squares method, and the model was used for forecasting monthly births and deaths.

2.2. Holt-Winters' Seasonal Method

The Holt-Winters method is a univariate automatic forecasting method that uses simple exponential smoothing (Holt 1957; Winters 1960). The forecast is obtained as a weighted average of past observed values in which the weight function declines exponentially with time, i.e., recent observations contribute more to the forecast than earlier observations. Forecasted values are dependent on the level, slope and seasonal components of the series being forecast. The model-specific formulation depends on whether seasonality is modelled in an additive or multiplicative way. The additive method is selected when the seasonal variations are approximately constant through the series, whereas the multiplicative method is preferred when the seasonal variations change proportionally to the level of the series (Hyndman and Athanasopoulos, 2018). The additive method is specified as:

$$\begin{aligned} l_t &= \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} - b_{t-1}) \\ b_t &= \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \\ s_t &= \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m} \\ y_{t+h|t} &= l_t + hb_t + s_{t-m+h} \end{aligned} \quad (3)$$

where l_t, b_t and s_t denote the level, trend and seasonal components, respectively, with corresponding smoothing parameters α, β and γ; $y_{t+h|t}$ is the forecast for h periods ahead at time t. The Holt-Winters' **multiplicative method** is defined as:

$$\begin{aligned} l_t &= \alpha\frac{y_t}{s_{t-m}} + (1-\alpha)(l_{t-1} - b_{t-1}) \\ b_t &= \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \\ s_t &= \gamma\frac{y_t}{(l_{t-1} + b_{t-1})} + (1-\gamma)s_{t-m} \\ y_{t+h|t} &= (l_t + hb_t)s_{t-m+h} \end{aligned} \quad (4)$$

We initialize the model's hyperparameters using the decomposition approach suggested by Hyndman et al. (2008) and implemented in the forecast package in R. After examining each time series for both the

additive and multiplicative versions of the Holt-Winters' seasonal method, we finally selected the model showing lower residual sum of squares to produce forecasts of monthly births and deaths.

2.3. Exponential smoothing state space model

State Space models consist of a measurement equation that describes the observed data, and some state equations that describe how the unobserved components or states (level, trend, seasonal) change over time. The general Gaussian state space model involves a measurement equation relating the observed data to an unobserved state vector $x_t = (b_t, s_t, s_{t-1}, \dots, s_{t-(m-1)})$, an initial state distribution and a Markovian transition equation that describes the evolution of the state vector over time state. In this study, we investigate both the additive and multiplicative error versions of SS models that underlie the exponential smoothing methods of the form (Hyndman et al., 2002):

$$Y_t = \mu_t + k(x_{t-1})\varepsilon_t \quad (5)$$

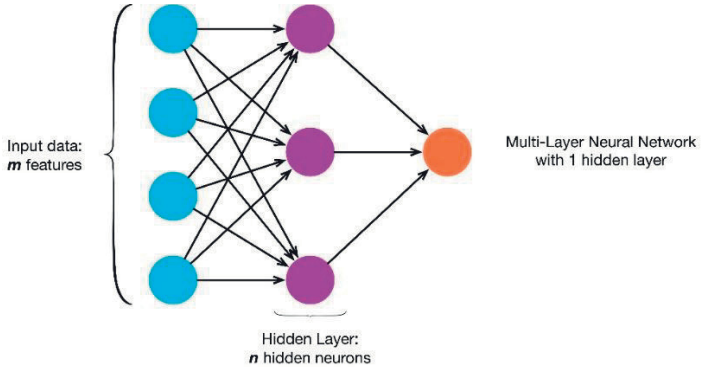
$$x_t = f(x_{t-1}) + g(x_{t-1})\varepsilon_t \quad (6)$$

where $\varepsilon_t \sim N(0, \sigma^2)$, $\mu_t = Y_{t-1}$ and where, for additive error models $k(x_{t-1}) = 1$, such that $Y_t = \mu_t + \varepsilon_t$, whereas for multiplicative error models $k(x_{t-1}) = \mu_t$ such that $Y_t = \mu_t(1 + \varepsilon_t)$. Model estimation involves measuring the unobservable state (prediction, filtering and smoothing) and estimating the unknown parameters using MLE methods.

2.4. Artificial Neural Network Algorithms

Artificial neural networks are forecasting methods (algorithms) that are based on individual, interconnected units called neurons that allow complex nonlinear relationships between the response variable and its predictors. The typical neural network architecture consists of a network of "neurons" organised in layers in which the predictors form the bottom layer, the forecasts form the top layer and there may be intermediate layers containing hidden neurons (Figure 1).

Figure 1 – A hypothetical example of Multilayer Perceptron Network.



Source: Author's preparation

In this latter case, each layer of nodes receives inputs from the previous layers, the neural network becomes non-linear and is known as a multilayer feed-forward network (MFFN). Forecasts are obtained by a linear combination of the inputs (or features) through an activation (or transfer) function, with weights automatically selected using a learning algorithm that minimises a cost function (Hyndman and Athanasopoulos, 2018). The following equation summarises the forecasted output.

$$f(\boldsymbol{x}, \boldsymbol{w}) = \varphi(\boldsymbol{x} \cdot \boldsymbol{w}) = \varphi\left(\sum_{j=1}^{P}(x_j w_j) + \theta\right) \quad (7)$$

where $\boldsymbol{x}$ and $\boldsymbol{w}$ represent, respectively, the input vector and weight vector of the neuron when there are P inputs into the neuron, θ is a bias term and φ denotes an activation function (e.g., sigmoid function). The process results in a single output from a neuron. With time series data, lagged values of the time series are used as inputs to a neural network in what is called a neural network autoregression. In this study we only consider MFFN with one hidden layer and add the last observed values from the same month as inputs.

2.5. Bootstrapp Aggregating (Bagging)

Bootstrap aggregating or bagging is a machine learning ensemble meta-algorithm that involves generating m new training sets by sampling from the original time series $D = \{(t_1, y_1), \ldots, (t_k, y_k)\}$ uniformly and with replacement. Then, for each of the m bootstrap samples, time series methods are fitted and combined using ensemble techniques. More formally,

Figure 2: The Bagging algorithm

Bagging Algorithm

Input: Training set $D = \{(t_1, y_1), \ldots, (t_k, y_k)\}$
B number of iterations
Forecasting methods / Base learners M

Process:

1. **for** $t = 1, \ldots, M$:
   $t_t = \varphi(D, D_{bs})$ % D_{bs} is the bootstrap distribution
   $Apply\ method\ M$ to the dataset D_{bs} and obtain $f^{bs}(x)$
2. ***Return***: $ensemble\ \{f^1(x), \ldots, f^B(x)\}$
3. **Prediction** with ensemble

$$f(\boldsymbol{x}) = \frac{1}{B}\sum_{h=1}^{B} f^h(x)$$

Source: Author's preparation

The idea is that a set of individual potentially weak learners (models) can be combined (averaged) to create a strong learner that outperforms individual models by reducing variance, and bias. First, the time series is Box-Cox-transformed, and then decomposed into trend, seasonal and residual components. We then bootstrapp the residual series, add them back to the trend and seasonal components, and reverse the Box-Cox transformation to obtain variations on the original time series. In this study we use 1000 bootstrapped series in combination with an SS model.

3 Research Methodology

3.1. Research Design

We set out a backtesting framework applicable to single-period ahead forecasts with steps:[3] (i) Selection the metric of interest (monthly births or deaths by sex and subpopulation); (ii) For each time series and period, selection of the historical "lookback window" for model calibration. We adopt a time series cross-validation approach, i.e., we consider a multi-step forecasting approach with re-estimation in which the training data or base period is extended before re-selecting and re-estimating the model at each iteration and computing forecasts. We adopt an expanding lookback window approach; (iii) Selection of the forecasting horizon ("lookforward window") over which to make forecasts. We focus on short-term horizon forecasts (1-year ahead of monthly births and deaths forecasts, i.e.,12 observations) since our interest is to use them as an input for computing MERP; (iv) Select a rolling fixed-length horizon backtesting approach in which we consider the accuracy of forecasts over fixed-length horizons as the jump-off date moves sequentially forward through time; (v) Select the evaluation criteria. We computed several criteria but, due to space constrains, we report the results for the Mean Absolute Percent Error (MAPE). For a given lookback and lookforward window, the MAPE for model j is defined as

$$MAPE_j = \frac{1}{n}\sum_{t=1}^{n}\frac{\left|\hat{y}_{t,j} - y_t\right|}{y_t} \times 100 \qquad (8)$$

where n is the number of forecasted values, $\hat{y}_t$ is the number of monthly births/deaths predicted by the model for time point t, and y_t is the corresponding value observed at time point t.

[3] For a similar approach used in evaluating the forecasting performance of stochastic mortality models and interest rate and credit risk models see, e.g., Dowd et al. (2010), Bravo & Silva (2006) and Chamboko & Bravo (2016, 2019a,b).

Each of the different time series models constructed (using a different lookback window and jump-off year) implies a different set of prediction intervals for the forecast horizon. To better understand the performance of the models analysed in terms of predicting the uncertainty of future births and deaths we computed the number of birth and death counts falling outside the 95% prediction intervals associated with each set of forecasts. Parameter estimation and model forecasting assessment were carried out using a computer routine written in R-script (R Development Core Team 2019).

3.2. Data

In this study, we use demographic data for Portugal comprising monthly data on live births and deaths broken down by sex and 25 different NUTS 3 regions from January 2000 to December 2018 provided by Statistics Portugal. The demographic dataset consists of 228 monthly observations for each one of the 100 different subpopulations of different size, the smallest with 38,753 resident individuals in December 2017 (Beira Baixa, male), the largest with 1,505,435 individuals (Lisbon Metropolitan Area, female). Of the 100 subpopulations tested, four (Lisbon and Oporto metropolitan areas male and female populations) correspond to highly populated areas with, in the case of Lisbon, more than one million residents. In contrast, the dataset tested includes several small population areas with less than 50,000 residents (e.g., Beira Baixa, Alto Tâmega, Alentejo Litoral). This archive is a challenging dataset in which to assess the monthly forecasting performance of time series methods since the data exhibits significant trend and seasonal components and high volatility in some cases, particularly in small population areas. Figure 2 represents the time series plot of monthly deaths of one representative small NUTS3 subpopulation.

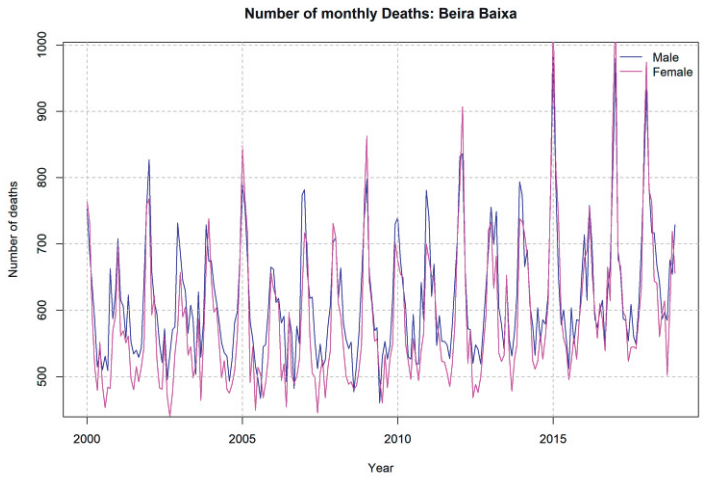


Figure 2 – Number of monthly deaths: Beira Baixa NUTS3 Region

4 Empirical results

The five time series methods are used as predictive models for making forecasts for future values of live births and deaths by sex and NUTS3 regions in Portugal. The MAPE results of 1-year ahead forecasts of monthly births and deaths by sex and NUTS3 regions for the period 2014-2018 averaged over all jump-off years with the different models are given in Tables 1 and 2, respectively. The results averaged (simple and weighted averages) over all 25 regions and five launch years are shown in the Tables. Additionally, Tables 1 and 2 include data on the population size of each NUTS3 region in December 2017 to ascertain whether the model's relative forecasting performance is a function of population size.

We first discuss the results related to monthly births forecasting. The all regions and launch years simple and weighted average forecasting performance for the five methods tested are similar for both male and female subpopulations showing relatively small average MAPE results, with the exception of ANN with one hidden layer that clearly underperforms. The simple average results show that the precision of the SARIMA forecasts is better than that of HW, SS and Bagging methods for the female subpopulations whereas, for the male counterparts, the Bagging method exhibits slightly lower forecasting errors.

Table 1 – Births Forecasting - Average MAPE by Model, Sex and NUTS3

Births	Females						Males					
NUTS 3	Popul.	AR	HW	SS	ANN	Bag	Popul.	AR	HW	SS	ANN	Bag
Alto Minho	124583	13.20	13.38	14.13	15.69	13.60	107595	12.46	13.07	12.87	12.55	12.61
Cávado	211950	10.23	10.39	9.63	12.70	9.65	192003	11.61	10.44	9.62	10.83	10.06
Ave	215975	10.50	9.57	8.67	11.71	8.80	197879	9.52	6.90	8.13	8.19	7.08
Área Metrop. Porto	910200	6.14	5.35	5.45	7.75	5.48	809502	4.74	5.40	5.04	5.58	5.16
Alto Tâmega	46044	18.49	18.86	21.03	19.17	20.38	41113	23.66	21.11	21.26	26.71	21.95
Tâmega e Sousa	216999	8.96	8.39	9.18	9.86	8.93	201769	8.62	7.61	8.55	8.17	8.51
Douro	101142	13.49	13.76	13.97	14.54	12.95	90904	14.27	14.86	13.88	15.12	13.31
Terras Trás-os-Montes	56870	15.33	15.02	16.95	18.77	15.75	51677	16.38	16.44	15.70	19.94	16.30
Oeste	186405	9.91	10.18	9.53	10.52	9.52	171301	7.84	8.82	8.07	11.46	7.95
Região de Aveiro	190926	8.84	8.75	8.22	10.11	8.19	172169	8.47	7.56	6.95	7.73	7.33
Região de Coimbra	231654	8.15	8.21	7.84	10.48	8.00	205294	7.86	7.70	7.34	9.88	7.16
Região de Leiria	149784	9.20	8.85	7.84	9.80	8.06	136525	9.86	10.79	9.75	10.78	9.83
Viseu Dão Lafões	134679	12.15	11.62	12.21	12.91	12.31	119952	12.64	12.53	12.12	13.38	11.85
Beira Baixa	43061	21.60	21.31	24.19	26.73	24.07	38753	16.40	17.92	17.23	18.77	16.70
Médio Tejo	123699	10.53	11.54	12.01	10.81	11.70	110956	9.99	10.63	10.21	14.13	10.59
Beiras, Serra da Estrela	114163	12.72	12.43	12.97	13.17	12.14	102025	11.11	10.75	12.37	11.08	11.73
Área Metrop. Lisboa	1505435	3.38	3.60	3.11	5.07	3.09	1328244	3.59	4.18	3.29	6.89	3.41
Alentejo Litoral	47551	16.99	17.56	17.78	18.75	17.68	46223	17.32	18.77	19.11	18.76	18.13
Baixo Alentejo	60669	11.59	12.45	12.57	12.41	12.33	57199	14.33	14.17	14.59	16.94	14.84
Lezíria do Tejo	124049	8.00	9.50	8.66	10.01	8.73	114666	11.56	10.94	11.48	12.16	11.08
Alto Alentejo	56092	18.80	19.01	18.54	18.80	18.39	50965	16.32	16.33	17.15	19.30	16.70
Alentejo Central	80677	12.81	13.87	13.01	14.05	13.05	73859	12.01	13.61	12.80	11.87	12.88
Algarve	229719	7.33	8.30	7.02	9.21	7.14	209898	7.40	7.46	7.24	8.45	7.13
RA Açores	125052	10.77	10.49	10.59	11.10	10.54	118810	9.78	9.78	9.52	9.53	9.27
RA Madeira	135957	12.00	12.30	14.02	13.22	12.97	118411	10.39	11.51	11.93	14.20	10.88
All regi.Simple Average	216933	11.64	11.79	11.96	13.09	11.74	194708	11.53	11.57	11.45	12.90	11.30
Weighted Average		8.00	7.99	7.83	9.49	7.74		7.70	7.87	7.52	9.30	7.48
Max	1505435	21.60	21.31	24.19	26.73	24.07	1328244	23.66	21.11	21.26	26.71	21.95
Min	43061	3.38	3.60	3.11	5.07	3.09	38753	3.59	4.18	3.29	5.58	3.41

Source: Authors preparation; **Notes:** Average Mean Absolute Percent Error (MAPE) by model (AR=ARIMA; HW; SS, ANN, Bagging) Sex and NUTS3 Region for the period 2014-2018. Weighted Average computed using the proportion of region's male or female population in the corresponding (sex) total population.

Note, however, that when considering the weighted average results (with weights given by the proportion of the region's subpopulation in the total resident population) the bagging method in combination with an exponential smoothing state space model exhibit higher forecasting accuracy due to their superior performance in highly populated regions. Using this later metric, the Bagging model advantages the SARIMA, HW, SS and ANN models by 0.26 (0.22), 0.25 (0.39), 0.09 (0.04) and 1.76 (1.82) percentage points in the female (male) subpopulations, respectively. On average for all models and for 59.2% of the subpopulations the forecasting errors are smaller for the male subpopulations when compared to their female counterparts. As expected, the average MAPE results over the five launch years are larger, the smaller the region's population size. The largest average forecasting error (26.73%) is found in the Beira Baixa female subpopulation using the ANN model whereas the highest accuracy (having 3.09% MAPE) is attained in the Lisbon metropolitan area ("Área Metropolitana de Lisboa") using the Bagging model. The forecasting error is less than 10% in 52.8% of the subpopulations considered.

Moving now to the results related to 1-year ahead monthly deaths forecasting, Table 2 shows once again that, with the exception of ANN that underperforms, the all regions and launch years simple and weighted average forecasting performance for the different models was relatively similar for both the male and female subpopulations, although the differences between the worst and the best performing model is higher in the male subset. Compared to births results, the average (weighted) forecasting accuracy of the alternative univariate time series methods continue to be lower in the male subpopulations and higher in the female group. The weighted average results show that the precision of SARIMA forecasts is consistently better than that of the HW, SS, ANN and Bagging models although the differences towards the predictive performance of SS and Bagging methods is small. The SARIMA model advantages the HW, SS, ANN and Bagging models by 0.59 (0.31), 0.19 (0.09), 1.51 (1.35) and 0.10 (0.11) percentage points in the female (male) subpopulations, respectively. On average for all

models and for 71.2% of the subpopulations the forecasting errors are notably smaller for the male subpopulations when compared to their female counterparts.

Table 2 – Deaths Forecasting - MAPE by Model, Sex and NUTS3

Births	Females						Males					
NUTS 3	Popul.	AR	HW	SS	ANN	Bag	Popul.	AR	HW	SS	ANN	Bag
Alto Minho	124583	10.36	11.27	10.64	10.74	10.59	107595	9.20	9.29	9.44	12.48	9.41
Cávado	211950	11.45	11.21	11.07	12.28	11.05	192003	8.88	9.14	8.99	9.03	8.84
Ave	215975	7.73	9.27	8.55	8.57	8.69	197879	8.89	9.31	9.05	10.79	8.92
Área Metrop. Porto	910200	7.24	8.07	7.85	8.87	7.79	809502	5.76	6.33	6.12	7.99	6.10
Alto Tâmega	46044	11.71	12.00	11.35	13.68	11.43	41113	12.69	15.07	14.94	15.14	14.73
Tâmega e Sousa	216999	9.02	11.21	10.33	13.00	10.05	201769	8.94	9.61	8.99	9.96	9.02
Douro	101142	11.25	13.74	12.21	13.57	12.27	90904	9.88	10.27	9.73	11.51	9.29
Terras Trás-os-Montes	56870	11.46	12.35	11.35	12.32	11.37	51677	10.53	11.07	10.71	12.72	10.59
Oeste	186405	7.30	8.11	7.65	9.07	7.73	171301	8.09	7.99	7.75	9.22	7.75
Região de Aveiro	190926	10.29	10.47	10.04	9.98	9.99	172169	7.94	9.20	8.20	9.33	8.34
Região de Coimbra	231654	7.54	7.56	7.57	7.44	7.25	205294	7.29	7.16	7.38	8.03	7.45
Região de Leiria	149784	9.57	9.98	9.63	10.92	9.68	136525	9.74	9.90	9.62	10.28	9.65
Viseu Dão Lafões	134679	9.91	10.35	9.80	12.78	9.88	119952	8.28	8.79	7.80	10.53	7.80
Beira Baixa	43061	14.26	14.13	14.96	14.22	14.95	38753	12.75	11.73	13.95	13.97	13.06
Médio Tejo	123699	8.10	7.95	7.74	9.56	7.57	110956	8.87	9.12	9.11	9.77	8.87
Beiras, Serra da Estrela	114163	10.29	11.48	10.34	11.36	10.42	102025	8.46	8.10	8.07	9.44	8.07
Área Metrop. Lisboa	1505435	6.01	6.07	5.89	7.70	5.71	1328244	5.07	5.01	4.99	5.59	5.16
Alentejo Litoral	47551	11.97	13.04	11.46	12.76	11.39	46223	13.24	16.11	15.24	15.97	14.86
Baixo Alentejo	60669	11.80	13.06	12.48	13.91	12.38	57199	10.09	10.29	10.00	12.26	9.99
Lezíria do Tejo	124049	9.48	10.33	9.97	11.11	10.05	114666	9.07	9.85	8.87	11.62	8.96
Alto Alentejo	56092	10.65	11.56	10.57	13.54	10.52	50965	11.29	11.71	11.48	13.08	11.50
Alentejo Central	80677	9.74	10.49	10.93	11.35	10.39	73859	9.22	9.52	8.98	11.82	9.25
Algarve	229719	9.26	9.65	8.94	10.88	8.96	209898	7.57	7.50	7.33	8.84	7.32
RA Açores	125052	10.90	11.72	11.33	11.94	11.07	118810	9.67	11.31	10.52	12.52	10.75
RA Madeira	135957	9.78	10.86	9.82	11.45	9.77	118411	9.53	10.05	9.50	11.17	9.54
All regi..Simple Average	216933	9.88	10.64	10.10	11.32	10.04	194708	9.24	9.74	9.47	10.92	9.41
Weighted Average		8.25	8.83	8.44	9.76	8.35		7.35	7.66	7.43	8.70	7.46
Max	1505435	14.26	14.13	14.96	14.22	14.95	1328244	13.24	16.11	15.24	15.97	14.86
Min	43061	6.01	6.07	5.89	7.44	5.71	38753	5.07	5.01	4.99	5.59	5.16

Source: Authors preparation; **Notes:** Average Mean Absolute Percent Error (MAPE) by model (AR=ARIMA; HW; SS, ANN, Bagg) Sex and NUTS3 Region for the period 2014-2018. Weighted Average computed using the proportion of region's male or female population in the corresponding (sex) total population.

Similar to the births results, the average MAPE results over the five launch years are smaller, the more populated the region is. The largest average forecasting error (16.11%) is found in the Alentejo Litoral male subpopulation using the HW model whereas the highest accuracy (4.99%) is attained in the Lisbon metropolitan area ("Área Metropolitana de Lisboa") male subpopulation using the SS model. The forecasting error is less than 10% in 37.6% of the subpopulations considered. To measure how well models perform in terms of predicting the uncertainty of future monthly birth/death counts over 1-year forecasting horizons, we computed the percentage of monthly birth and death counts falling outside the 95% prediction interval estimated for each model, sex and NUTS3 Region.[4] The results show that for the birth and death count forecasting exercises the prediction intervals for the SARIMA, SS and Bagging models consistently provide appropriate measures of uncertainty for short-term forecasting horizons. The SARIMA, SS and Bagging models perform equally well in terms of predicting the uncertainty of future monthly death counts, with SS and Bagging models slightly overperforming in births forecasting. On the contrary, the HW and ANN model consistently fail in predicting the uncertainty of future monthly birth and deaths with, in some regions, up to 17% of observed death counts falling out of the 95% prediction interval (Alto Minho, ANN model).

[4] Due to space constraints, the full set of results is not displayed in the study but can be obtained from the authors upon request.

5 Conclusion

Monthly time series of live births and deaths exhibit significant and persistent seasonality patterns, requiring the adoption of appropriate forecasting methods to increase the accuracy of population forecasts. In this study we empirically evaluated the forecasting performance of traditional time series methods (seasonal ARIMA, HW and SS) and new machine learning approaches applied to birth and death monthly forecasting by sex and NUTS 3 regions for Portugal using a backtesting framework and monthly data for the period 2000-2018. With the exception of the ANN structure with a single hidden layer that clearly underperformed, the all regions and launch years simple and weighted average forecasting performance for the three models was relatively similar for both male and female subpopulations births and deaths. However, our results show that the Bagging method combined with an exponential smoothing state space model exhibit higher forecasting accuracy for births whereas for deaths forecasting the seasonal ARIMA slightly overperformed both alternative univariate time series methods and machine learning algorithms. As expected, the weighted average precision is higher, the more populated the region is. The prediction intervals for the SARIMA, SS and Bagging models consistently provide appropriate measures of uncertainty for short-term forecasting horizons. Further research should check for the robustness of these results against alternative forecasting horizons and fixed lookback windows using rolling fixed-length horizon backtests. Future research will also investigate the robustness of these results against alternative primary, extended, composite, and hybrid performance metrics used in machine learning regression, forecasting and prognostics, considering for competing distance measures and normalization and aggregation procedures.

References

Abel, G., Bijak, J., Forster, J., Raymer, J., Smith, P. & Wong, J., (2013) Integrating uncertainty in time series population forecasts: An illustration using a simple projection model. *Demographic Research*, 29 (43), 1187-1226.

Alho, J., Bravo, J. M. & Palmer, E. (2013). *Annuities and Life Expectancy in NDC*. In Holzmann, R. E. Palmer and D. Robalino (Eds.), Nonfinancial defined contribution Pension Schemes in a Changing Pension World, Vol. 2 Gender, Politics, and Financial Stability, 395 - 436.

Ayuso, M., Bravo, J. M. & Holzmann, R. (2020). Getting Life Expectancy Estimates Right for Pension Policy: Period versus Cohort Approach. Journal of Pension Economics and Finance, 1-20. doi: doi:10.1017/S1474747220000050;

Ayuso, M., Bravo, J. M., & Holzmann, R. (2017a). Addressing Longevity' Heterogeneity in Pension Scheme Design. *Journal of Finance and Economics*, 6(1), 1–21.

Ayuso, M., Bravo, J. M., & Holzmann, R. (2017b). On the Heterogeneity in Longevity among Socioeconomic Groups: Scope, Trends, and Implications for Earnings-Related Pension Schemes. *Global Journal of Human Social Sciences - Economics*, 17(1), 31–57.

Booth, H. (2006). Demographic forecasting: 1980 to 2005 in review. International Journal of Forecasting, 22 (3), 547-581.

Bravo, J. M. (2016). Taxation of Pensions in Portugal: A Semi-Dual Income Tax System. *CESifo DICE Report - Journal for Institutional Comparisons*. 14 (1), 14-23.

Bravo, J. M. (2019). Funding for Longer Lives: Retirement Wallet and Risk-Sharing Annuities. EKONOMIAZ *Basque Economic Review*, Nº 96 (II-2019), 268–291.

Bravo, J. M. & Coelho, E. (2019). Forecasting Subnational Demographic Data using Seasonal Time Series Methods. In *Proceedings of the 19th Portuguese Association of Information Systems Conference: digital disruption: living between data science, IoT and ... people* (pp. 40). Associação Portuguesa de Sistemas de Informação. ISSN 2183-489X

Bravo, J. M. & Coelho, E. (2020). Forecasting small population monthly fertility and mortality data with seasonal time series methods. In: Linhares, W. (Ed.). As Ciências Sociais Aplicadas e a Interface com vários Saberes 2, Atena Editora, 158-176.

Bravo, J. M. & Coelho, E. (2020). Modelling monthly birth and deaths using Seasonal Forecasting Methods as an input for population estimates. In: *Demography of Population Health, Aging and Health Expenditures. The Springer Series on Demographic Methods and Population Analysis*. In Press.

Bravo, J. M. & Silva, C. (2006). Immunization Using a Stochastic Process Independent Multifactor Model: The Portuguese Experience. *Journal of Banking and Finance*, 30 (1), 133-156.

Bravo, J. M., & El Mekkaoui de Freitas, N. (2018). Valuation of longevity-linked life annuities. *Insurance: Mathematics and Economics*, 78, 212–229.

Bravo, J. M., & Herce, J. A. (2020). Career breaks, Broken pensions? Long-run effects of early and late-career unemployment spells on pension entitlements. Preprint submitted to Journal of Pension Economics and Finance.

Bravo, J. M., Ayuso, M., Holzmann, R. & Palmer, E. (2020). Addressing Life Expectancy Gap in Pension Policy. Preprint submitted to Insurance: Mathematics and Economics.

Bravo, J. M., Coelho, E., & Magalhães, M. G. (2010). *Mortality projections in Portugal*. In EUROSTAT - European Commission (eds.), Work session on demographic projections, EUROSTAT-EC Collection: Methodologies and working papers, Theme: Population and Social Conditions, 241–252.

Bravo, J. M., Rodrigues, T., Ribeiro, S. & Inácio, A. (2018). *Portugal. Projeções de População Residente 2011-2040*. In Teresa Rodrigues & Marco Painho (Coord.). Modelos Preditivos e Segurança Pública. Fronteira do Caos Editores, 168-214.

Chamboko, R., & Bravo, J. M. (2016). On the modelling of prognosis from delinquency to normal performance on retail consumer loans. *Risk Management*, 18(4), 264–287.

Chamboko, R. & Bravo, J. M. (2019a). Modelling and forecasting recurrent recovery events on consumer loans. *International Journal of Applied Decision Sciences*, 12(3), 271-287.

Chamboko, R. & Bravo, J. M. (2019b). Frailty correlated default on retail consumer loans in developing markets. *International Journal of Applied Decision Sciences*, 12(3), 257–270.

Dowd, K., Cairns, A., Blake, D., Coughlan, G. Epstein, D. & Khalaf-Allah, M. (2010) Backtesting Stochastic Mortality Models, North American Actuarial Journal, 14:3, 281-298

EU (2016). The impact of demographic change on European regions. European Union - Committee of the Regions (doi:10.2863/26932).

Holt, C.C. (1957). Forecasting seasonals and trends by exponentially weighted averages. O.N.R. Memorandum 52/1957, Carnegie Institute of Technology. Reprinted with discussion in 2004, *International Journal of Forecasting*, 20, 5–13.

Hyndman, R. & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. 2nd edition, OTexts.

Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, 18, 439–454.

Hyndman, R. J., Koehler, A.B., Ord, J. K., & Snyder, R.D. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*. Springer.

Keilman, N., Pham, D., & Hetland, A. (2002). Why population forecasts should be probabilistic – illustrated by the case of Norway. *Demographic Research*, 6, 409–453.

Lee, R. (1992). Stochastic demographic forecasting. *International Journal of Forecasting*, 8, 315–327.

Lee, R., & Tuljapurkar, S. (1994). Stochastic population forecasts for the United States: Beyond high, medium, and low. *Journal of the American Statistical Association*, 89, 1175–1189

Li S., Yang, Z., Li, H. & Shu, G. (2018). Projection of population structure in China using least squares support vector machine in conjunction with a Leslie matrix model. Journal of Forecasting, 37(2), 225–34.

Pflaumer, P. (1992). Forecasting U.S. population totals with the Box–Jenkins approach. *International Journal of Forecasting*, 8, 329–338.

R Development Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL www.R-project.org.

Tayman, J., Smith, S. K., and Lin, J. (2007). Precision, bias, and uncertainty for state population forecasts: an exploratory analysis of time series models. *Population Research and Policy Review,* 26(3), 347–369.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, 6, 324–342

Wiśniowski A., Smith P., Bijak, J., Raymer, J. & Forster, J. (2015). Bayesian Population Forecasting: Extending the Lee-Carter Method. Demography. Jun; 52(3), 1035-59.

Support Vector Machine para Imputação e Edição de Valores
O caso das Declarações Mensais de Remuneração das Empresas

Filipe Santos, *filipe.santos@ine.pt*
Departamento de Metodologia e Sistemas de Informação, Instituto Nacional de Estatística

Pedro Campos, *pedro.campos@ine.pt*
Departamento de Metodologia e Sistemas de Informação, Instituto Nacional de Estatística &
Universidade do Porto/Faculdade de Economia

1. Introdução

Na recolha de dados através de questionários, alguns dos inquiridos podem não responder a uma ou mais questões. Nessas situações ocorre o que se designa por dados em falta, ou "não repostas". Pode também acontecer que algumas das repostas tenham sido mal recolhidas, ou introduzidas de forma incompleta ou errada. Em todos estes casos é habitual recorrer-se a métodos de edição e de imputação de valores. Durante os processos de edição e imputação são estimados novos valores para substituir os valores errados ou para imputar os valores em falta. O tratamento de valores em falta (que em inglês, de forma geral, se designam como *nonresponse* ou *missing* data) constitui um problema clássico em Estatística, estudado por vários autores (Schafer, 1997, Little e Rubin, 2002, De Waal et al., 2011, entre outros).
O INE divulga trimestralmente, estatísticas sobre remunerações, com base na informação da Declaração Mensal de Remunerações transmitidas pelas empresas à Segurança Social e da Relação Contributiva dos subscritores da Caixa Geral de Aposentações. Pretende-se com este trabalho descrever o processo de imputação das Declarações Mensais de Remuneração (DMR) provenientes da Segurança Social (SS), informação que o INE tem tratado como forma de estimar os indicadores da remuneração bruta mensal média[1]. Além destes indicadores, o INE utiliza a mesma fonte para atualizar outros inquéritos, tais como o IVNE (Índice de Volume de Negócios e Emprego) e o ICTE (Índice do Custo de Trabalho - Empresas). Em concreto, pretende-se com este trabalho corrigir ou imputar valores das variáveis associadas às remunerações, embora também se considerem outras variáveis para imputação, nomeadamente o número de trabalhadores, e contribuições. Para a deteção das empresas que corrigem significativamente a informação serão utilizados dois processos: (i) um critério, denominado *ad hoc*, que contabiliza o número de correções significativas durante o ano anterior e (ii) outro critério baseado em Support Vector Machine (SVM). A informação em causa diz respeito a cerca de 400 mil empresas e a aproximadamente 4,2 milhões de trabalhadores. Pretende-se, no futuro, quando o INE dispuser de informação ao nível do trabalhador, poder complementar estas estatísticas por outras que permitam, entre outras possibilidades, conhecer a distribuição das remunerações e proceder à caracterização sociodemográfica dos trabalhadores. O artigo encontra-se estruturado da seguinte forma: no Capítulo 2 começa-se por fazer uma descrição geral dos métodos de imputação e de edição, incluindo alguma revisão de literatura e aplicações no INE. No Capítulo 3 descreve-se o processo de imputação e edição das Declarações Mensais de Rendimentos (DMR). No capítulo 4 apresentam-se os resultados e conclusões.

2. Métodos de imputação e de edição

2.1. Breve revisão de literatura

De acordo com Waal et al, (2011), existem dois tipos de métodos de imputação: simples e múltipla. No primeiro caso é calculado um valor para substituir o valor em falta. Na imputação múltipla usa-se mais

[1] Parte desta informação encontra-se disponível em "Statslab" do portal do INE:
https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_statslab

do que um valor. Existe ainda a imputação multivariada, em que o objetivo é fazer imputação de valores de variáveis diferentes para o mesmo indivíduo. Os processos de edição e de imputação podem ser modelizados de seguinte forma. Seja Y^k uma matiz de dados (n x q) com n respostas e q variáveis relativas à unidade de tempo k, que inclui as variáveis observadas Y_j^k, com j=1,2,…q, e os dados em falta. Cada indivíduo i, (em geral, um pessoa individual ou empresa), responde através de um vetor de respostas completo que, no instante k, pode ser representado da seguinte forma:

$$Y_{ij}^k = \left(Y_{i1}^k, Y_{i2}^k, \dots, Y_{iq}^k\right)$$

Seja R a matriz indicatriz dos casos i em falta relativamente à variável j. A matriz R pode ser definida da seguinte forma no instante k:

$$R_{ij}^k = \begin{cases} 1, se\ Y_{ij}^k\ está\ em\ falta\ no\ instante\ k \\ 0, se\ Y_{ij}^k\ é\ observado\ no\ instante\ k \end{cases}$$

Seja S a matriz indicatriz dos casos i a editar relativamente à variável j. A matriz R pode ser definida da seguinte forma no instante k:

$$S_{ij}^k = \begin{cases} 1, se\ Y_{ij}^k\ deve\ ser\ editado\ no\ instante\ k \\ 0, se\ Y_{ij}^k\ não\ deve\ ser\ editado\ \ no\ instante\ k \end{cases}$$

As matrizes R e S são binárias e correspondem, na verdade, aos valores de Y. Os processos de imputação simples podem ser mais elementares, tais como a imputação pela média, imputação por regressão, ou usando algoritmos de *machine learning*, tais como Support Vector Machine, ou Redes Neuronais e ainda imputação hot-deck. Nos métodos hot-deck, é imputado um valor em falta a partir de um registo semelhante selecionado aleatoriamente. Nos métodos de regressão e nos outros métodos de machine learning são usadas variáveis auxiliares para prever o valor em falta na variável objetivo. Também existem outros métodos de imputação simples, tais como Expectation Maximisation (EM) e o Método da Máxima Verosimilhança. Finalmente, quanto aos mecanismos de não respostas, podemos ainda acrescentar que estes podem ser classificados como MCAR (Missing Completely At Random), MAR (Missing At Random) e NMAR (Not Missing At Random). No caso MCAR, a probabilidade de que um valor esteja em falta não depende da variável objetivo (variável a imputar) nem das variáveis auxiliares. Admite-se que a distribuição das não respostas a uma determinada variável é idêntica à distribuição das respostas à mesma variável. No caso MAR, a probabilidade de um valor estar em falta depende das variáveis auxiliares, mas não da variável a imputar. No caso NMAR, a probabilidade de um valor estar em falta depende das variáveis auxiliares e também da variável a imputar. Esta última situação pode acontecer, por exemplo, nos inquéritos que contenham variáveis de rendimento das famílias, nos casos em que os rendimentos são tendencialmente altos. No caso presente, o mecanismo de não resposta é NMAR (Non Missing At Random), uma vez que os valores em falta dependem das variáveis auxiliares e também da variável a imputar. Também no caso presente, uma vez que se consideram para imputação outras variáveis, (nomeadamente o número de trabalhadores, remunerações e contribuições), estamos, assim, perante um processo de imputação simples multivariada, porque apenas se usa um valor para imputar o valor em falta, embora se imputem valores de variáveis diferentes para o mesmo indivíduo.

2.2. Estimação de não-respostas no INE

Em geral, o INE segue a metodologia proposta por Harrell (2001), segundo o qual se a proporção de não respostas for inferior ou igual a 5% pode ser adotada a imputação simples; se a proporção de não respostas se situar entre os 5% e os 15% é sugerido o uso de imputação múltipla. Os métodos de tratamento de não respostas de várias operações estatísticas[2] do INE consideram imputação simples, enquanto o método para o tratamento de não respostas do IUTICE[3], por exemplo, envolve imputação múltipla. Nos casos de imputação simples a imputação de não respostas é feita, em geral, apenas para os não respondentes do ano n (ano de referência dos dados) que respondem no ano anterior (n-1). No caso do IUTICE, a metodologia é aplicada a todas as unidades inquiridas que não tenham respondido à

[2] São exemplos as seguintes operações: IMPA (Inquérito aos Municípios - Proteção do Ambiente), IONGA (Inquérito às Organizações Não Governamentais do Ambiente), IEDCB (Inquérito às Entidades Detentores de Corpos de Bombeiros) e ISBSA (Inquérito ao Setor de Bens e Serviços de Ambiente)

[3] IUTICE – Inquérito à Utilização das Tecnologias de Informação e Comunicação nas Empresas

variável *compras*. Neste caso a metodologia de imputação tem por base a informação disponível para os anos n-2, n-1 e n, do volume de negócios e da Informação Empresarial Simplificada (IES) relativa às variáveis que dão origem à variável compras. As regras segundo as quais a variável *compras* é imputada dependem da atividade económica em causa e as imputações são efetuadas com base em variáveis como o *Custo das mercadorias vendidas* e das *matérias consumidas* e *Fornecimentos e serviços externos*, Gastos administrativos e Fornecimentos e serviços externos.

3. Imputação e edição das Declarações Mensais de Rendimentos (DMR)

3.1. Considerações gerais sobre os dados

Os dados provenientes da Segurança Social (SS) são recebidos pelo INE mensalmente, sendo referentes às remunerações declaradas pelas empresas nos meses anteriores. Para cada mês k são entregues quatro versões de informação. A existência de diferentes versões resulta do facto das empresas poderem corrigir *a posteriori* a informação anteriormente declarada. Acontece que para meses diferentes do trimestre se recorre a versões diferentes do histórico na fase de deteção das empresas a imputar. Essas versões corrigidas são referidas mais à frente como sendo as versões k-1, k-2, k-3 e k-4, sendo k o mês de referência. Das vinte e cinco variáveis que compõem os registos (microdados) destacam-se para efeitos de imputação treze variáveis[4]. Nos dados da Segurança Social existem valores em falta (*missings*) relativos às remunerações para algumas das empresas. De modo a poder responder às necessidades de diferentes operações estatísticas do INE, foi acordado identificar e imputar as empresas não respondentes (*missings*) assim como identificar antecipadamente as empresas que farão alterações significativas nas DMR nas versões seguintes do valor reportado. O processo de imputação visa preencher valores omissos que se admitem como sendo não nulos. Para este efeito considera-se como significativa uma alteração no valor que ultrapasse os dez mil euros. Foram testados outros possíveis valores, mas este limiar foi o que permitiu obter uma taxa de cobertura de correções elevada sem comprometer o processo de deteção de empresas a imputar. Além disso, este valor tem associado um baixo número de falsos positivos.

3.2. Metodologia de seleção

São imputados os dados dos meses k-1, k-2 e k-3, sendo o primeiro aquele que é sujeito a um maior número de imputações. De referir que para a seleção das empresas a imputar, tem-se em conta aquelas que corrigem significativamente a informação. Serão utilizados dois processos: (i) um critério que denominamos como *ad hoc*, que seleciona as empresas a imputar tendo por base o número de correções significativas durante o ano anterior e outro (ii) que seleciona as empresas através em algoritmos de *machine learning* recorrendo ao *SVM (Support Vector Machine).* No critério *ad hoc* para identificação dos casos a imputar, as empresas com correções significativas para um mês *k* do trimestre a imputar são identificadas através da construção de um conjunto de dados com o histórico de correções para o ano anterior a *k*. Tal conjunto é obtido agregando os ficheiros dos doze meses anteriores a *k* construídos no passo anterior e respeitando a versão associada ao mês *k*. Por exemplo, para o segundo mês do trimestre são consideradas as correções a partir da segunda versão. De acordo com o critério *ad hoc* são selecionadas para imputação todas as empresas contendo:

- Pelo menos 9 meses de correções significativas para o ano anterior;
- Pelo menos 3 meses de correções significativas para o quadrimestre anterior.

[4] A lista de variáveis com as respetivas descrições é demasiado extensa para ser incluída neste artigo. Como exemplo, refira-se a variável remuneração, cuja imputação depende da sua natureza. Para o salário base, bónus de carácter mensal, subsídios de caráter mensal, subsídio de refeição e trabalho noturno recorre-se ao valor apresentado na última versão do mês anterior. Para o subsídio de férias, de Natal, prémios, bónus e subsídios de caracter regular não mensal recorre-se ao valor homólogo respeitando a variação do salário base entre do mês k-12 e k-1. Para a compensação por cessação de contrato de trabalho recorre-se a: (i) mediana dos valores desde k-12 a k-1 caso se observem pelo menos seis observações não nulas nesse período; (ii) mínimo desde k-12 a k-1 no caso de se observarem 4 ou 5 observações não nulas; (iv) nas restantes situações o valor imputado é nulo. Para as restantes categorias da natureza de remuneração recorre-se à mediana dos valores desde k-12 a k-1 se existirem pelo menos seis observações não nulas durante esse período, caso contrário o valor imputado será igual ao apresentado na última versão do mês k-1.

3.3. Algoritmo SVM e o procedimento de imputação

O Support Vector Machine (SVM) (Hastie et. al, 2017) é um algoritmo de *machine learning* supervisionado que consiste num processo de classificação binário baseado em regressão. Os dados de treino são constituídos por N pares. O algoritmo procura encontrar o hiperplano que melhor se adequa à classificação dos dados. O hiperplano é construído de modo a maximizar a margem entre o mesmo e os pontos mais próximos, denominados pontos de suporte, que definem os vetores de suporte, d=+1 e d=-1.

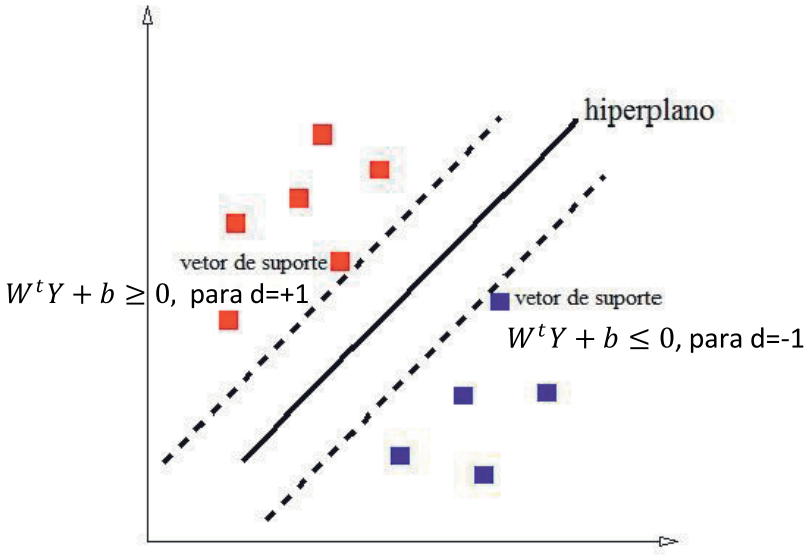


Fig. 1. Hiperplano e pontos de suporte no SVM

Este mesmo hiperplano permite a classificação dos dados relativos ao conjunto de teste. Para se poder implementar o SVM constroem-se para o mês *k* (a imputar) dois ficheiros: um de treino e outro de teste. O ficheiro de treino contém o histórico de correções para o mês *n-1*, ou seja, trata-se de um ficheiro com 13 colunas: as 12 primeiras contêm informação das correções por empresa para os 12 meses anteriores a *k-1* e a 13ª contém a identificação para o mês *k-1* das empresas que efetivamente corrigiram significativamente a declaração nesse mês (valores superiores a 10 mil euros). Ou seja, para *n-1* temos o histórico do ano anterior e respetiva identificação de correção. Todas as colunas contêm valores binários. Trata-se na verdade de um vetor binário $Y_{ij}^k \in \mathbb{R}^q$, com q=13, e sendo a variável objetivo dada por $Y_{ij}^k = \{0,1\}, \forall\, j = 1, \dots, q$. O valor $Y_{ij}^k = 0$ corresponde à ausência de edição ou imputação, enquanto que $Y_{ij}^k = 1$ corresponde às situações em que os valores foram editados. O ficheiro de teste que está relacionado com o mês *k* contém apenas o histórico do ano anterior, uma vez que ainda se desconhece se as empresas irão ou não corrigir a informação no mês *k*. Tal como para o critério *ad hoc,* este procedimento respeita a versão associada ao mês a imputar. Com recurso ao *package* Sklearn do *Python* (Buitinck et al., 2013), é aplicado o SVM ao conjunto de treino e, posteriormente, o modelo obtido é aplicado ao conjunto de teste para efeitos de validação. O processo de imputação decorre da seguinte forma:

a) São consideradas como empresas a imputar no mês k todas as empresas com mais de nove trabalhadores sem registos nesse mês e com pelo menos um registo não nulo no mês k-1. Ou seja, são consideradas empresas a imputar, (designadas por *i**), as empresas em que:

$$i^* = \left\{R_i^k = 1, \wedge\, \exists_j : R_i^{k-1} = 0, \forall j \in Y^k\right\}$$

b) São consideradas empresas a editar os seus resultados no mês k todas as empresas que foram selecionadas através dos seguintes processos:

b.1) Via método "inteligente", recorrendo ao mês k-1 e ao seu histórico do ano anterior para construir o conjunto de treino que posteriormente é implementado no mês k e no seu histórico do ano anterior. Ou seja, o fato de no mês k-1 termos informação se a empresa fez ou não correções significativas permite fazer projeções para o mês k (para o qual não temos essa informação).

b.2) Empresas com pelo menos 9 meses de correções significativas para o ano anterior; ou pelo menos 3 meses de correções significativas para o quadrimestre anterior. Ou seja,

$$i^* = \{ (S_{ij}^{k-1} = 1 \wedge S_{ij}^{k-2} = 1 \wedge \cdots \wedge S_{ij}^{k-9} = 1) \vee (S_{ij}^{k-1} = 1 \wedge S_{ij}^{k-2} = 1 \wedge S_{ij}^{k-2} = 1 \wedge S_{ij}^{k-3} = 1 \wedge S_{ij}^{k-4} = 1), \forall j \in Y^k \}$$

Seja o vetor de treino (Y_{ij}^k, Y_{ij}^{13}) , sendo $Y_{ij}^k \in \mathbb{R}^q$, com i=1,…, n empresas e j=1,…, q variáveis e k=n-1, …, n-12 instantes temporais. As q variáveis correspondem às características de entrada. O vetor de saída $W \in \mathbb{R}^q$ é constituído pelos pesos, uma para cada variável j=1,…,q, cuja combinação linear prevê $Y^k \in \mathbb{R}^q$. O objetivo do SVM é encontrar um vetor de pesos W , que permita definir um hiperplano com a forma: $W^t Y + b = 0$, em que b é um vetor de resíduos e W é um vetor de pesos e se verificam as seguintes condições: $W^t Y + b \geq 0$, para d=+1 ; $W^t Y + b \leq 0$, para d=-1, sendo d os vetores de suporte como assinalados na Fig.1 Está em causa um Lagrangiano que se pretende minimizar: $min\left(L_p\right) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} a_i y_i (x_i\, w + b\,) + \sum_{i=1}^{l} a_i$

4. Resultados

Na Tabela 1 apresenta-se a taxa de acerto do modelo SVM associado a uma subamostra de 10% do conjunto de treino (que contém cerca de 400 mil empresas) entre os meses de outubro de 2019 a fevereiro de 2020. O conjunto de teste é relativo às mesmas empresas e é relativo ao mês de março de 2020. Para cada mês são imputados dados até à terceira versão dos dados, sendo esta a versão que menos correções sofre. Em geral, na primeira versão são identificadas cerca de 500 empresas que fazem correções significativas; na segunda 80 unidades e na terceira apenas 40.

Mês a imputar	1ª Versão	2ª Versão	3ª Versão
Outubro	-	-	0.9999
Novembro	-	0.9998	0.9999
Dezembro	0.9995	0.9998	0.9999
Janeiro	0.9995	0.9998	0.9999
Fevereiro	0.9994	0.9997	-
Março	0.9994	-	-

Tabela 1: Taxa de acerto do SVM numa subamostra de 10% do conjunto de treino

Ano	Mês	Valor Inicial	Valor Pós Imputação	Valor Final	Variação Pós Imputação %	Variação Inicial %	Nº empresas
2017	10	1 388 829 371	1 459 660 276	1 477 055 511	-1.18	-5.97	10 599
2017	11	2 097 419 302	2 126 103 548	2 146 421 442	-0.95	-2.28	10 603
2017	12	1 677 589 881	1 773 231 134	1 806 331 284	-1.83	-7.13	10 604
2018	1	1 495 565 307	1 549 600 260	1 564 413 684	-0.95	-4.40	10 605
2018	2	1 394 475 085	1 482 254 916	1 481 294 084	0.06	-5.86	10 604
2018	3	1 310 883 954	1 506 299 781	1 526 912 508	-1.35	-14.15	10 604
2018	4	1 526 282 065	1 564 001 745	1 577 625 702	-0.86	-3.25	10 604
2018	5	1 521 946 393	1 611 636 727	1 633 314 917	-1.33	-6.82	10 605
2018	6	1 837 328 401	1 877 253 169	1 912 809 821	-1.86	-3.95	10 603
2018	7	1 705 410 946	1 784 315 282	1 795 488 000	-0.62	-5.02	10 603
2018	8	1 471 147 336	1 576 427 885	1 591 465 485	-0.94	-7.56	10 602
2018	9	1 467 163 844	1 516 916 674	1 528 263 225	-0.74	-4.00	10 603

Tabela 2: Valores e correspondentes variações Pós Imputação e Inicial face a valor Final (tomado como referência para efeitos de cálculo de taxas de variação)

Como referido acima, um dos objetivos deste trabalho é a utilização desta informação em vários inquéritos, nomeadamente no Índice de Volume de Negócios e Emprego (IVNE). Na tabela 2 registam-se as diferenças observadas para as cerca de 10600 empresas relativas ao IVNE entre outubro de 2017 e setembro de 2018. Registou-se o total de remunerações para três situações distintas:

1. dados da primeira versão da Segurança Social (**Valor Inicial**);
2. dados imputados da primeira versão após deteção de *missings* e empresas com correções significativas via *machine learning* (**Valor Pós Imputação**);
3. dados da última versão da Segurança Social (**Valor Final**).
   É calculada a variação entre as duas primeiras situações (**Variação Pós Imputação** e **Variação Inicial)** e a **Variação final**. Os ganhos obtidos com o processo de imputação parecem ser significativos para todos os meses considerados nesta análise.

5. Discussão

Face aos resultados obtidos parece evidente a melhoria relativamente à primeira versão dos dados provenientes da Segurança Social devida ao processo de imputação. Não só se antecipam possíveis correções como se identificam os casos de *missings* mais significativos. O trabalho de edição/imputação de valores das variáveis associadas às remunerações encontra-se ainda numa fase experimental. No entanto, os dados editados e imputados têm o potencial de utilização noutras operações estatísticas em que o INE utiliza a mesma fonte de dados, tais como no IVNE (Índice de Volume de Negócios e Emprego) e o ICTE (Índice do Custo de Trabalho - Empresas), sem comprometer os prazos definidos para a divulgação de resultados. Além disso este tipo de tratamento, baseado em algoritmos de *machine learning*, poderá permitir a criação de novos indicadores estatísticos relativos ao emprego e remunerações.

Referências

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., (and other authors) (2013), API design for machine learning software: experiences from the scikit-learn project. European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases, Sep 2013, Prague, Czech Republic.

De Waal, T., de Waal, Pannekoek, J., Scholtus, S., (2011), Handbook of Statistical Data Editing and Imputation, Wiley

Hastie, R. Tibshirani, and J. Friedman, (2017), The Elements of Statistical Learning, 2nd Ed., Springer Series in Statistics Springer New York Inc., New York, NY, USA, (2001)

Harrel, F., (2001), Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis, Springer

Little, R.J.A. and Rubin, D.B. (2002) Statistical Analysis with Missing Data, John Wiley & Sons, New York.

Schafer, J.L. (1997) Analysis of Incomplete Multivariate Data, Chapman & Hall, London.

O papel da intermediação do setor financeiro – quando o todo é menor do que a soma das partes

Filipa Lima, *slima@bportugal.pt*
Departamento de Estatística, Banco de Portugal

Sónia Mota, *scmota@bportugal.pt*
Departamento de Estatística, Banco de Portugal

Ângela Coelho, *afcoelho@bportugal.pt*
Departamento de Estatística, Banco de Portugal

1. Introdução

Tal como referido por Plašil e Kubicová (2012), a análise das ligações entre setores institucionais permite uma melhor compreensão do processo de contágio da economia, expondo os setores mais sensíveis. A crise financeira internacional agravada pela crise da dívida soberana alterou as relações intersectoriais e evidenciou a necessidade de novas ferramentas de análise sobre o impacto das decisões financeiras dos diversos setores.

Neste artigo é feita uma extensão ao trabalho de Lima e Monteiro (2014) sobre as interligações dos setores na economia portuguesa, na medida que, em vez de considerar o setor financeiro agregado, isola o papel de cada subsetor do setor financeiro. Os gráficos de fluxos de fundos, baseados nos dados das contas financeiras, são complementados com o Índice de Poder de Dispersão e o Índice de Sensibilidade de Dispersão, propostos por Tsujimura e Mizoshita (2004) e Okuma (2012).

2. Abordagem metodológica

De acordo com Bodie *et al.* (2010), a economia é vista como uma estrutura inter-relacionada de ativos, passivos e garantias entre diversos setores institucionais: as sociedades não financeiras (SNF), as sociedades financeiras (SF), as administrações públicas (AP), os particulares (Part) e o setor não residente (RM). As sociedades financeiras englobam diferentes tipos de entidades com funções financeiras muito distintas entre si: o banco central (BC), as outras instituições financeiras monetárias (OIFM[1]), os outros intermediários financeiros e auxiliares financeiros (OIFAF[2]) e as sociedades de seguros e fundo de pensões (SSFP). Analisar as sociedades financeiras de forma agregada ou isolar cada um destes tipos pode apontar para resultados distintos, em função do sentido e dimensão da contribuição de cada tipo para o valor agregado. A abordagem seguida é consistente com o sistema de contas financeiras.

As relações intersectoriais são analisadas nos anos de 2007 (período anterior à crise financeira internacional), 2011 (início do PAEF, programa de assistência económica e financeira a Portugal), 2015 (período após o final do PAEF) e 2018, (período pós-crise financeira), por via das transações líquidas de cada ano, ou seja, transações em ativos financeiros deduzidas de transações em passivos, dos vários instrumentos financeiros. O fluxo de fundos deve ser interpretado da seguinte forma: o diâmetro do

[1] As OIFM correspondem aos bancos, no essencial.
[2] Aqui incluem-se os fundos de investimento, os outros intermediários financeiros exceto sociedades de seguros e fundos de pensões, os auxiliares financeiros e as instituições financeiras cativas e prestamistas.

círculo é proporcional à poupança financeira de cada setor, sendo usada a cor cinza em caso de capacidade líquida de financiamento e a cor preta em caso de necessidade líquida de financiamento. As setas ilustram os fluxos financeiros líquidos do setor credor para o setor devedor e a espessura é proporcional à magnitude dessas relações. Sempre que o setor financeiro é financiador líquido de outro setor, a seta é representada a cinza e quando os outros setores financiam o setor financeiro, a seta é representada a preto. As relações que não envolvem o setor financeiro não serão representadas nos fluxos de fundos desagregados, para facilitar a leitura dos mesmos, no entanto, poderão ser observadas a tracejado nos fluxos de fundos agregados.

Recorrendo à relação matricial definida por Tsujimura e Mizoshita (2004), numa ótica de passivos (financiamentos):

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1S} \\ x_{21} & \ddots & & x_{2S} \\ \vdots & & \ddots & \vdots \\ x_{S1} & x_{S2} & \cdots & x_{SS} \end{bmatrix},$$

$$\sum_{j=1}^{S} x_{ij} = A_i \quad \text{e} \quad \sum_{i=1}^{S} x_{ij} = L_j$$

onde o elemento x_{ij} representa a magnitude da exposição entre o setor credor i e o setor devedor[3] j.
A soma A_i, da linha i, corresponde ao total de ativos do setor i face a todos os setores devedores e a soma L_j, da coluna j, corresponde ao total de passivos do setor j, face a todos os setores credores.

Adicionalmente, tal como definido por Tsujimura e Mizoshita (2004) e Okuma (2012), X pode ser usado para calcular o Índice de Poder de Dispersão (PDI, p_j) e o Índice de Sensibilidade de Dispersão (SDI, s_i). O PDI indica a influência que uma unidade de choque na procura de financiamento do setor j tem na procura de financiamento de outros setores. Por outro lado, o SDI indica a influência que uma unidade de choque na procura total de financiamento tem na procura de financiamento do setor i. Esses índices são definidos da seguinte maneira:

$$p_j = \frac{\sum_{i=1}^{S} \gamma_{ij}}{\frac{1}{S}\sum_{j=1}^{S}\sum_{i=1}^{S} \gamma_{ij}}, \; s_i = \frac{\sum_{j=1}^{S} \gamma_{ij}}{\frac{1}{S}\sum_{j=1}^{S}\sum_{i=1}^{S} \gamma_{ij}},$$

onde γ_{ij} são os elementos da matriz Γ:

$$\Gamma = (\mathrm{I} - \mathrm{C})^{-1} = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1S} \\ \gamma_{21} & \ddots & & \gamma_{2S} \\ \vdots & & \ddots & \vdots \\ \gamma_{S1} & \gamma_{S2} & \cdots & \gamma_{SS} \end{bmatrix}$$

onde C representa a matriz de coeficiente de entrada, cujos elementos são dados por:

$$c_{ij} = \frac{x_{ij}}{t_i}, \text{ e } t_i = \max(A_i, L_j).$$

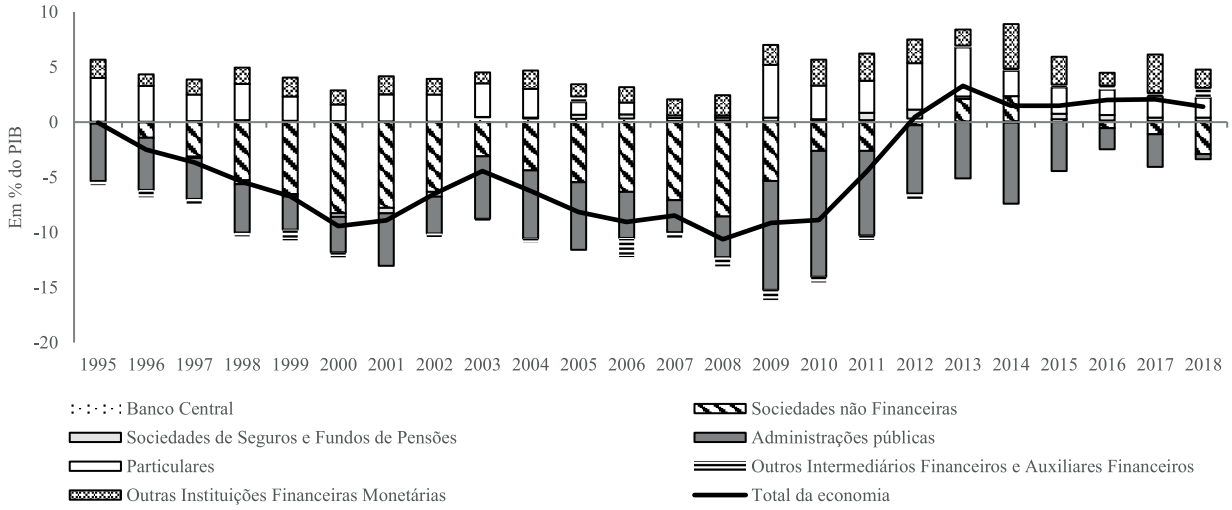
A matriz inversa indica a influência, direta e indireta, de uma mudança nos montantes de investimento de um setor (ativos) nos montantes de investimento de outros setores. A matriz X é construída a partir das estatísticas das contas financeiras, pelo facto de representar os ativos e passivos financeiros por setor credor / devedor.

[3] A relação entre dois setores não é necessariamente simétrica, pois, um setor pode ser credor de outro sem que ele seja seu devedor pelo mesmo montante.

3. Análise setorial – o caso de Portugal

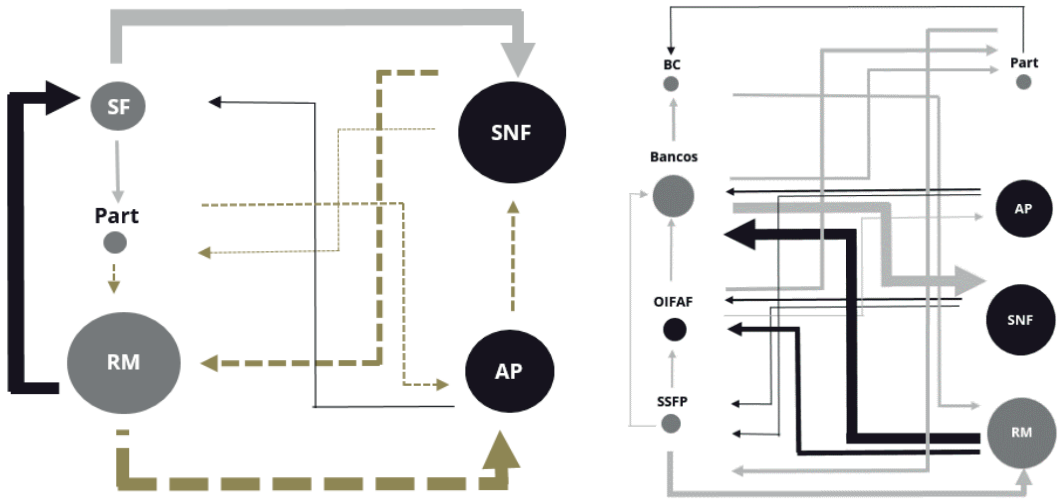
Até 2011, a economia portuguesa apresentou uma necessidade de financiamento face ao exterior, i.e., foi devedora líquida do RM (Gráfico 1), com um valor médio de 6.6 por cento do PIB. Em 2008, Portugal registou a maior necessidade de financiamento de 10.6 por cento do PIB, invertendo a tendência em 2012, ano a partir do qual passou a apresentar excedente externo.

Gráfico1 | Poupança financeira



Em 2007, a economia nacional apresentava uma necessidade de financiamento de 8.5 por cento do PIB. O RM era financiador líquido das AP e das SF, através da concessão líquida de empréstimos e da aquisição de títulos de dívida. O papel das SF era dominado pelos bancos que canalizavam fundos do RM para as SNF e para os particulares, através da concessão de empréstimos.

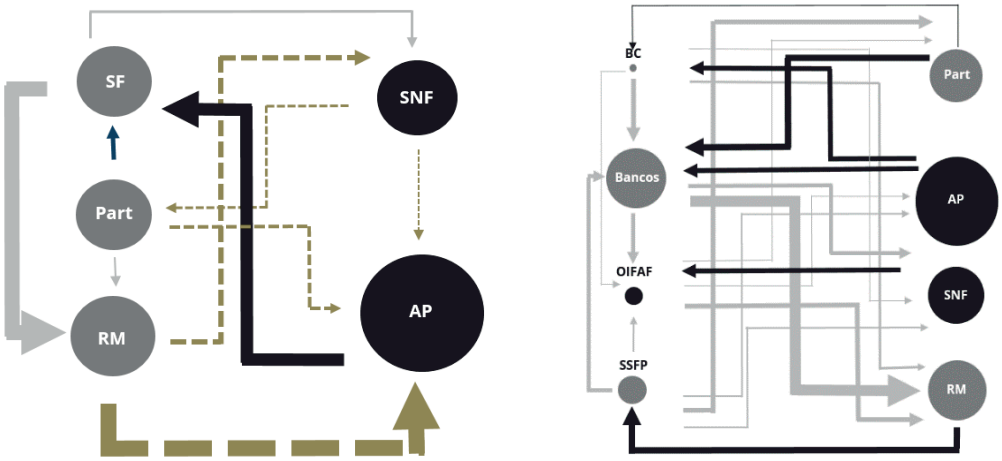
Gráficos 2 e 3 | Fluxo de fundos, 2007



Contudo, com o início do PAEF, as relações intersectoriais alteraram-se. Em particular, o sentido e dimensão das setas que ligam as SF às AP e ao RM variam consoante o tipo de SF analisada. As necessidades de financiamento da economia diminuíram para 4.5 por cento do PIB, com melhorias na poupança financeira de todos os setores, à exceção das AP. Apesar de continuarem a ser maioritariamente financiadas pelo RM, num contexto de difícil acesso aos mercados financeiros, a estrutura de financiamento das AP altera-se significativamente, apresentando uma contração do financiamento titulado e um aumento dos empréstimos externos. As AP surgem como financiadoras líquidas do SF o que resulta da colocação junto do BC dos montantes não utilizados dos fundos obtidos no PAEF e do apoio estatal concedido aos bancos. As SNF apresentam-se como o segundo setor com

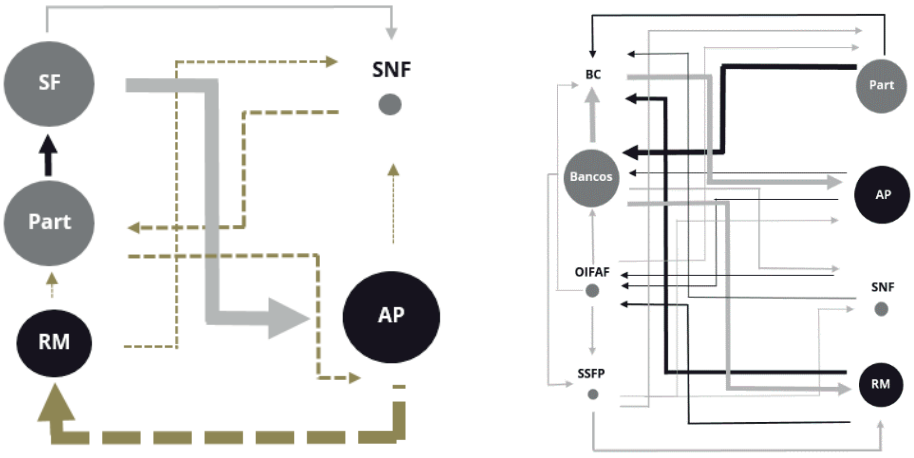
maiores necessidades de financiamento, alterando a intensidade das suas relações com o SF, nomeadamente com os bancos e os OIFAF. O reembolso líquido de empréstimos contribuiu para a redução do financiamento líquido dos bancos às SNF. Com a crise financeira, os particulares não só financiaram, em termos líquidos, as AP, como também os bancos. Esta nova realidade foi determinada pela amortização líquida de empréstimos e pelo desinvestimento em provisões técnicas de seguros, o que contribuiu para a alteração do padrão de financiamento entre as SSFP e o RM (não visível quando apenas se consideram as SF de forma agregada), pelo facto de as SSFP investirem uma parte substancial da sua carteira em títulos de dívida estrangeiros.

Gráficos 4 e 5 | Fluxo de fundos, 2011



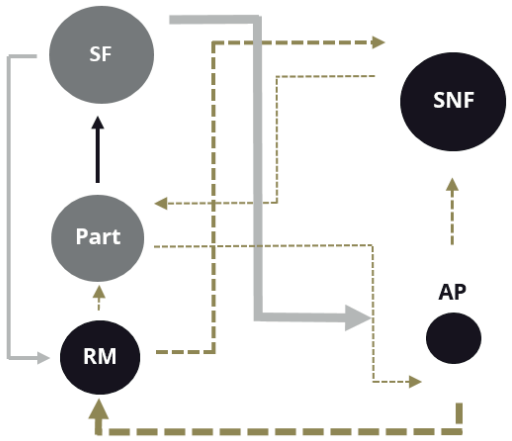
O ano de 2015 marca uma alteração significativa no sentido e na magnitude dos fluxos financeiros entre os vários setores, com a economia a apresentar uma capacidade de financiamento de 1.5 por cento do PIB. Com a implementação de medidas de política monetária não convencionais, o BC passa a ter um papel de destaque no SF, financiando (em termos líquidos) as AP, através da aquisição de títulos de dívida no âmbito do *public sector purchase programme* (PSPP). Adicionalmente, com o recurso a um fluxo de fundos mais detalhado, as operações entre o BC e os bancos tornam-se visíveis, destacando-se, neste ano, a redução das operações de cedência de liquidez. Num contexto de desalavancagem prevista no PAEF, as SNF passam a ter capacidade de financiamento e os particulares continuam a registar uma necessidade de financiamento líquido face aos bancos.

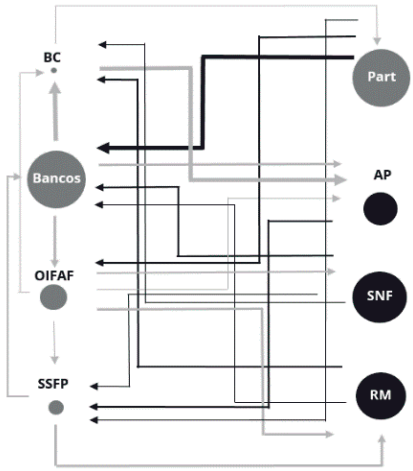
Gráficos 6 e 7 | Fluxo de fundos, 2015



A partir de 2015, as AP reduzem a sua necessidade de financiamento, amortizando parte dos empréstimos obtidos no âmbito do PAEF, parcialmente compensado pelo financiamento líquido do BC no âmbito do PSPP e do investimento dos particulares em títulos de dívida pública e nos certificados de aforro e do Tesouro.

Gráficos 8 e 9 | Fluxo de fundos, 2018

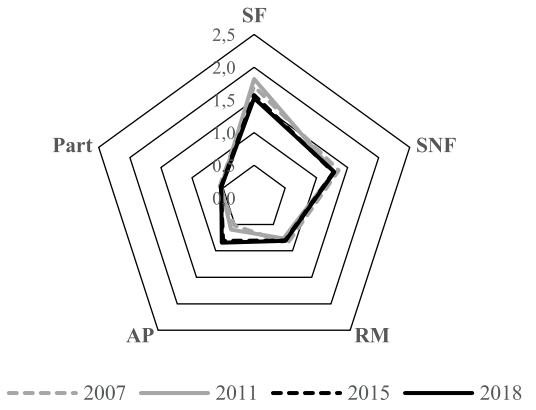


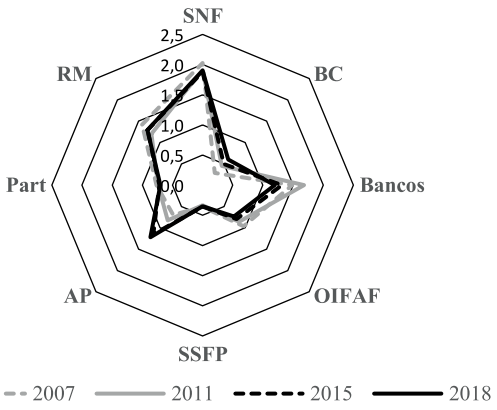


4. Análise do Índice de Poder de Dispersão e do Índice de Sensibilidade de Dispersão

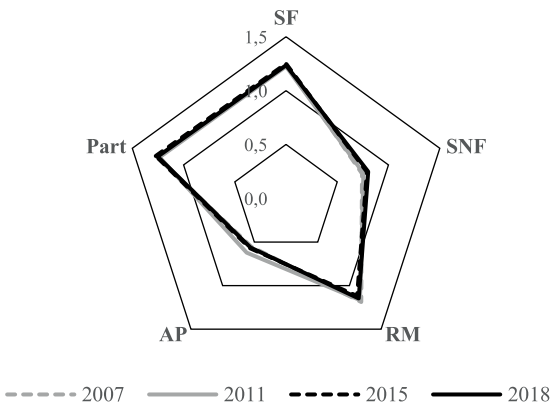
Os gráficos 10 e 13 representam o Índice de Poder de Dispersão (PDI) e o Índice de Sensibilidade de Dispersão (SDI), conforme proposto por Tsujimura e Mizoshita (2004) e Okuma (2012). A análise ao PDI e ao SDI entre os anos de 2007 e 2018 confirmam as conclusões de Tsujimura e Mizoshita (2004): *"The most prominent thing is that, the location of the plots on the diagram show minimal change despite the laps of time"*. Apesar da crise financeira e da crise da dívida soberana, os PDI e SDI mantiveram-se praticamente inalterados.

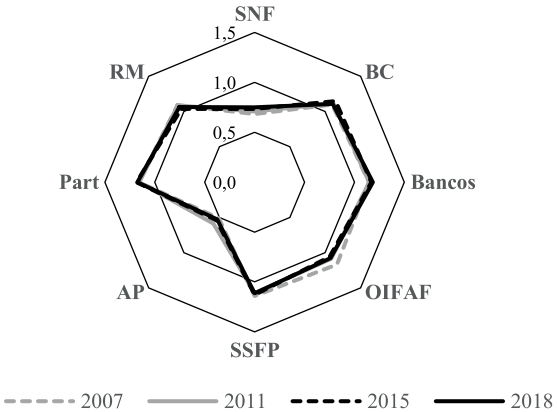
Gráficos 10 e 11 | PDI





Gráficos 12 e 13 | SDI





O PDI dos bancos e das SNF revela a importância dos dois setores na geração de financiamento na economia, em que os bancos confirmam o seu papel de intermediário financeiro com um PDI e SDI superior a um e as SNF pelo facto de atuarem como intermediário financeiro para as empresas do grupo (Lima e Monteiro, 2014). Como resultado do apoio ao sistema financeiro, a partir de 2015, as AP apresentam um PDI superior a um e um SDI inferior a um. Contudo, apenas supriram parte das necessidades de financiamento deste setor. No caso do SDI, é relevante referir o papel do RM, dos particulares e de todos os subsetores do SF, na capacidade de suprir necessidades de financiamento de outros setores. Do SF, destaca-se o papel do BC na relação com o sistema financeiro, mais especificamente aos bancos residentes, através do recurso às operações de refinanciamento de prazo alargado e às AP, através do PSPP. Quer no caso do PDI quer no SDI, os valores obtidos para o agregado das SF são diferentes dos valores individuais obtidos para o banco central, bancos e restantes subsetores.

5. Observações finais

O artigo destaca a importância do uso de dados mais desagregados para melhor entender realidades díspares dentro do próprio setor financeiro. A análise e o desenvolvimento de modelos relativos à interligação entre os setores são fundamentais, de modo a entender como os choques num setor se podem propagar nos outros setores. Para o efeito, contas integradas, fluxos de fundos e "quem-a-quem" são uma condição indispensável.

Agradecimentos

As autoras agradecem à Lídia Brás, Maria Teresa Crespo, Pedro Alves e Sérgio Branco pelos comentários e sugestões. As análises, opiniões e conclusões aqui expressas são da exclusiva responsabilidade das autoras e não refletem necessariamente as opiniões do Banco de Portugal ou do Eurosistema.

Referências

Bodie, Z., Gray, D. F. e Merton, R. C., 2010, "Measuring and Managing Macrofinancial Risk and Financial Stability: A New Framework", Central Bank of Chile, Studies, Volume 15: Financial Stability, Monetary Policy, and Central Banking.

Lima, F. e Monteiro, O., 2014, "How do macro-financial linkages adjust in times of adjustment? – evidence from Portugal". Supplement to the Statistical Bulletin, March 2016. Banco de Portugal.

Plašil, M. e Kubicová, I., 2012, "Contingent Claim Analysis and the Inter-Sector Transmission of Credit Risk", Czech National Bank / Financial Stability Report 2011/2012.

Tsujimura, K. e Mizoshita, M., 2004, "Compilation and application of asset-liabilities matrices: a flow-of-funds analysis of the Japanese economy 1954-1999," K.E.O Discussion Paper, No. 93.

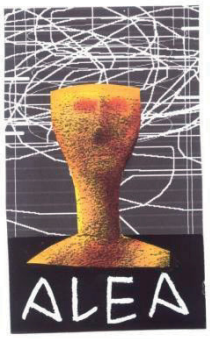
O INE como (pro)motor da Literacia Estatística

Francisco Correia, *francisco.correia@ine.pt*

Serviço de Difusão, Instituto Nacional de Estatística

«Para apreendermos o sentido de um texto, não basta sabermos soletrar as palavras que o constituem, é necessário utilizarmos um vasto conjunto de outros recursos. Do mesmo modo, para "lermos" informação estatística precisamos de conhecer conceitos, terminologias e metodologias usados na sua elaboração.»

in www.alea.pt, Apresentação



O INE aposta, há muito, em iniciativas que visam a promoção da literacia estatística em diversos contextos, consciente da importância que este saber assume numa sociedade evoluída. A primeira e porventura a mais emblemática dessas iniciativas foi o ALEA – Ação Local de Estatística Aplicada, nascido há mais de vinte anos de uma parceria com a Escola Secundária de Thomaz Pelayo[1].

O ALEA foi criado com o propósito de proporcionar instrumentos relacionados com a compreensão, a utilização e o ensino da Estatística, destinados essencialmente aos docentes e alunos do ensino secundário; mas afirma-se também como um importante meio de apoio a projetos interdisciplinares, do qual podem beneficiar outros públicos.

A sua ação é exercida fundamentalmente através de um sítio na Internet (www.alea.pt), que disponibiliza conteúdos de vária ordem no âmbito da Estatística – cursos "Noções de Estatística", "Noções de Probabilidades", "Inferência" e "Organização e Tratamento de Dados", dossiês pedagógicos, fichas de trabalho para uso em sala de aula (ActivAleas), Glossário, Desafios… – e da informação estatística – Estatísticas em Foco, EuropAlea, Países lusófonos, GeoEscolas… –, a que se acrescenta uma vertente lúdica: jogos didáticos, humor estatístico…

O ALEA tem sido apontado, à escala internacional, como exemplo de boas práticas. Em consequência, foi objeto da primeira atribuição do prémio "*Best Cooperative Project Award*" pela *IASE – International Association for Statistical Education*, em 2007.

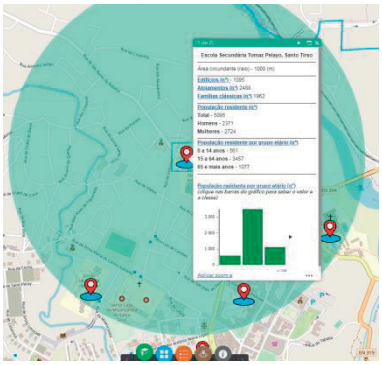
Outra iniciativa marcante foi a criação, em 2004, da Rede de Informação do INE em Bibliotecas do Ensino Superior (RIIBES), que em 2018 deu lugar à Rede de Informação do INE para o Ensino Superior (RIIES)[2].

No âmbito destas Redes, foram dinamizadas por técnicos do INE, desde 2010, centena e meia de sessões de divulgação/formação em instituições de ensino superior (IES) de todos os distritos de Portugal

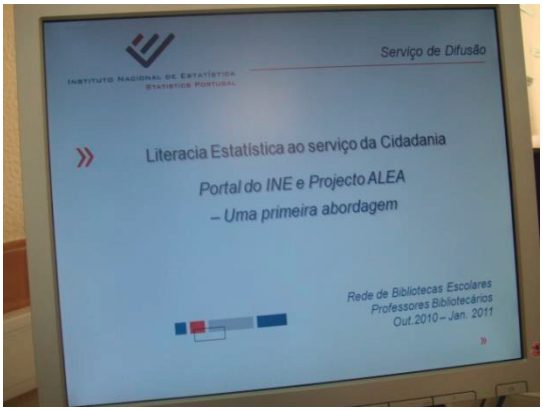
[1] Na qual se integrou depois a Direção Regional de Educação, que mais tarde deu lugar à Direção Geral de Estabelecimentos Escolares.

[2] A RIIBES – Rede de Informação do INE em Bibliotecas do Ensino Superior funcionou entre 2004 e 2017, relativamente a mais de 30 instituições de ensino superior com as quais foram estabelecidos protocolos. A RIIES tem uma dinâmica idêntica, mas sem a necessidade de protocolos.

continental, focadas no Portal do INE e no Portal do Eurostat, dirigidas a docentes, discentes e investigadores. Paralelamente, e desde o início, o INE ofereceu a bibliotecas de IES mais de mil publicações estatísticas.



Atividade idêntica surgiu em 2010, no âmbito de um protocolo estabelecido com o Gabinete da Rede de Bibliotecas Escolares. Neste contexto, realizaram-se cerca de 500 sessões de divulgação/formação sobre o Portal do INE e o ALEA em estabelecimentos escolares, dirigidas a professores de qualquer disciplina do ensino básico e secundário, também dinamizadas por técnicos do INE. Paralelamente, têm sido oferecidos, em cada ano, mais de 1000 exemplares do Anuário Estatístico de Portugal destinados a bibliotecas escolares.



Os seminários "Portas abertas" são mais uma iniciativa do INE no âmbito da promoção da literacia estatística. Constam igualmente de sessões de divulgação/formação, neste caso realizadas nas instalações do INE (Lisboa e Porto) e para o público em geral. Existem "Portas abertas" sobre o Portal do INE, o Portal do Eurostat ou um tema específico (p. ex.: Índice de Preços no Consumidor).

Seminários Portas Abertas – INE

Portal de Estatísticas Oficias
18 de outubro de 2019 | 10h00 – 12h00 | Lisboa e Porto
Gartuitos – Úteis para todos – Presenciais – Curta duração
Em Lisboa e no Porto

Esta Sessão, aberta gratuitamente a todos os interessados, vai dar a conhecer os inúmeros recursos e funcionalidades do Portal do INE, que oferece milhares de indicadores estatísticos sobre os principais temas da sociedade portuguesa, sempre acompanhados de metainformação para ajudar à leitura dos dados; Destaques; publicações; dossiês temáticos; estudos; aplicações interativas, etc. Além disso, vai mostrar-lhe como adaptar a informação existente às suas necessidades, através das opções e ferramentas disponíveis para obter, por exemplo, informação mais desagregada, ou com outro arco temporal. E, muito importante, de forma simples e rápida!
Esta sessão vai abrir as portas a um conhecimento evolutivo mais profundo, fiável e atual sobre o seu País, a sua região, o seu município e a sua freguesia!

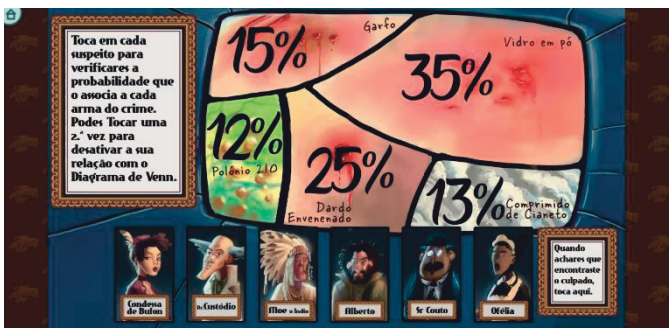
Noutra vertente, está a ser ultimada a Explorística 2.0, uma evolução da "Explorística – Aventuras na Estatística", criada em 2013 pela Sociedade Portuguesa de Estatística com o apoio da Ciência Viva e cujos direitos foram gentilmente cedidos ao INE. A Explorística 2.0 integra vários módulos interativos, envolvendo atividades diversas, tais como selecionar, recolher, descrever e estimar.

Estará disponível como exposição física, mais transportável, e nas seguintes versões virtuais:

– *online* via *browser*, com acesso livre e sem instalação de *software* adicional;

– *mobile* para Android e IOS.

Existe ainda a versão em realidade virtual de um dos módulos (Submarino).



No ano letivo 2017-2018, o INE integrou mais uma atividade na sua ação global no domínio da promoção da literacia estatística, ao participar na 1.ª Competição Europeia de Estatística – ESC2018. Tratou-se de uma iniciativa conjunta do Eurostat e vários Institutos Nacionais de Estatística, que foi bem-sucedida, teve continuidade e vai na sua 3.ª edição.

A ESC2020, tem a participação de equipas (alunos + tutor) de 17 países da União Europeia, distribuídas por duas categorias: A – ensino secundário e B – 3.º ciclo do ensino básico.

A Competição Europeia de Estatística tem duas fases:

1. A fase nacional, que este ano decorre entre janeiro e julho, durante a qual os alunos competem com colegas do seu país, realizando primeiro testes *on line* sobre consulta de informação estatística e interpretação de publicações estatísticas e depois apresentações (PPoint) a partir de uma base de dados estatísticos. A fase nacional da ESC2020 é, pela primeira vez, uma organização conjunta do INE e do Banco de Portugal;

2. A fase europeia, na qual os melhores classificados da fase nacional em cada categoria representam os seus países a nível europeu e produzem pequenos vídeos (2 minutos) com base num mesmo tema estatístico, que são avaliados por um júri internacional.

Em Portugal, o nível de inscrições na ESC tem crescido ano após ano:

	Alunos	Equipas
2018	564	206
2019	660	240
2020	1107	406

Contamos com os leitores deste Boletim como divulgadores da Competição Europeia de Estatística, para que assim continue a acontecer nas próximas edições da Competição!

INE: preservar a memória institucional

Paula Marques, *paula.marques@ine.pt*

Serviço de Difusão, Instituto Nacional de Estatística

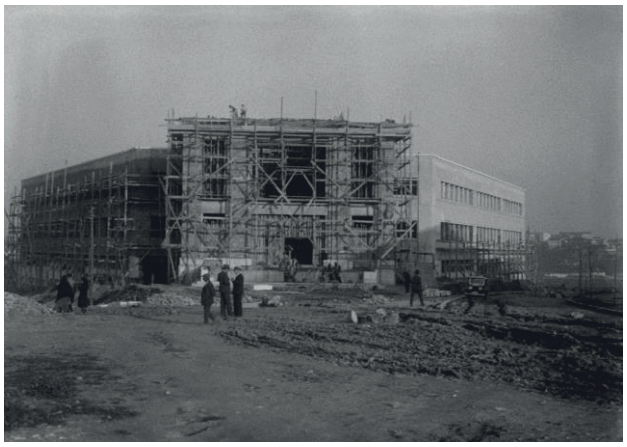
Preservar a memória institucional é criar um continuum entre passado e presente visando o futuro. O INE tem-se empenhado em conservar esse passado – materializado em objetos, fotografias, livros e documentos – dando-o a conhecer de diversas formas, pois confia que preservar a memória é uma espécie de diálogo entre tempos e fortalece a identidade institucional.
Este texto aborda duas temáticas relacionadas com a preservação da memória institucional: uma é dedicada ao edifício-sede e a outra é um vislumbre do património bibliográfico.

"O Palácio da Estatística"

Foi deste modo amplificado que Fernando Santiago se referiu ao edifício do Instituto Nacional de Estatística, num artigo publicado na revista Panorama em outubro de 1942. O jornalista não poupa elogios ao Edifício da Estatística e é generoso em epítetos: majestoso – um dos mais belos documentos da arquitetura contemporânea – um dos mais perfeitos da Europa.
Os aplausos estendem-se ao Instituto – enquanto organismo modelar – e aos seus dirigentes. Apesar de ser um artigo com laivos panegíricos, Fernando Santiago transmite ao leitor uma história vívida do edifício e do Instituto.
Graças a este artigo conseguimos reconstituir a vida do INE por dentro e por fora: a distribuição dos serviços, como se trabalhava no Palácio da Estatística, as obras de arte existentes, etc. Essas palavras permitem-nos "escutar" uma história que caiu no silêncio do tempo, após a partida de sucessivas gerações dos que trabalharam na casa do INE – é essa história vívida, escrita em 1942, que queremos partilhar com os leitores de 2020.
As imagens apresentadas pertencem ao Banco de Imagens do INE com exceção de três, devidamente identificadas. As legendas entre aspas e itálico são transcrições do artigo supracitado; as notas entre parenteses retos são comentários nossos para melhor compreensão das imagens ou complementação da informação.

"Linhas modernas – superfícies lisas, linhas direitas, portas e janelas largas e uma entrada monumental"

"Transposta a entrada, de linhas sóbrias e elegantes, encontra-se um átrio com duzentos metros quadrados onde se veem quatro grandes colunas de mármore de Pero Pinheiro"



"A sua construção foi iniciada há uma dezena de anos, mercê do espírito organizador do Sr. Eng.º Duarte Pacheco, ilustre ministro das Obras Públicas, sendo o projeto definitivo da autoria do arquiteto Pardal Monteiro"

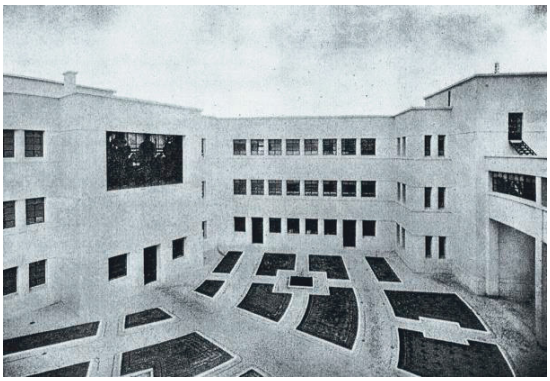
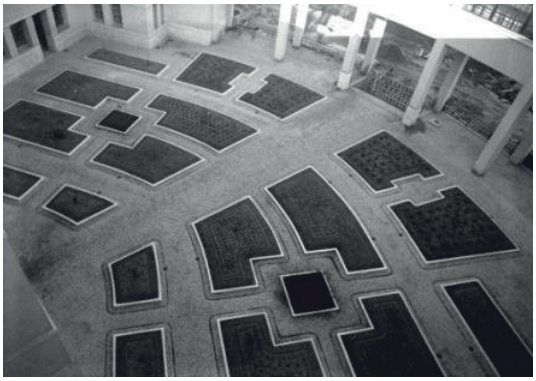
"Foi concluída no curto espaço de dois anos. A abertura dos caboucos começou em fevereiro de 1932, a 23 desse mês procedeu-se à cerimónia do lançamento da primeira pedra"

"(..) preparava-se (…) a respetiva laje-cobertura, a cujo enchimento se procedeu no dia 1 de fevereiro de 1933, isto é, precisamente um ano depois de iniciados os trabalhos" («O INE», ed. 1936)

[Nota: neste ato estiveram presentes os três membros da Comissão administrativa de acompanhamento da obra de construção do edifício que eram quadros dirigentes da Direção Geral de Estatística: Júlio Rangel de Lima, Casimiro Chambica da Fonseca e Artur Pena Martins]

"O Palácio da Estatística ocupa uma vasta área na vizinhança do Instituto Superior Técnico, outra obra notável que se deve à iniciativa do Sr. Eng.º Duarte Pacheco, e tem a forma de um A com três grandes pavimentos"

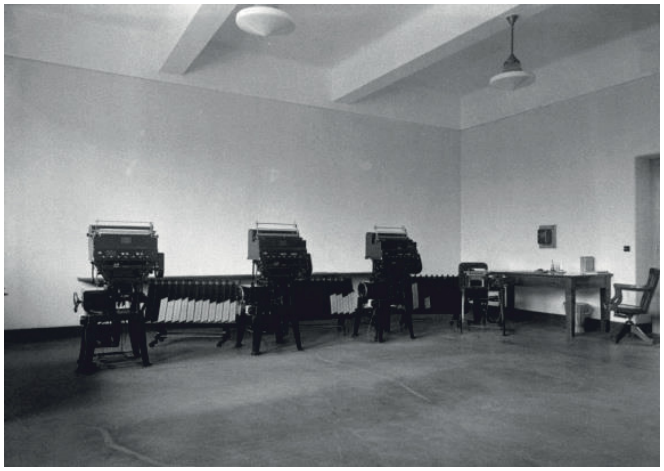

"(...) sendo o segundo e o terceiro[pavimentos] ligados por uma larga e vistosa galeria envidraçada que liga as duas alas"

(**) 


"O centro do edifício é embelezado por um grande páteo ajardinado"

"(...) dois anos depois ficaram instalados no novo edifício todos os serviços de estatística ocupando 56 dependências. E logo se verificou um maior rendimento de trabalho cuja perfeição tem sido objeto de rasgados louvores de nacionais e estrangeiros, considerando o nosso Instituto como um dos mais perfeitos da Europa"

[Nota: em notícia publicada pelo jornal Diário de Notícias, de 2 agosto de 1933, ficamos a saber que foi nessa data que foi efetuada a transferência do pessoal e dos serviços da Direção Geral de Estatística (DGE) para o novo edifício. Recorde-se que só em maio de 1935 será publicada a lei que cria o INE e extingue a DGE]



(**)

"No primeiro pavimento, ao norte, encontram-se as espaçosas salas de arquivo, e ao sul veem-se as casas das máquinas dos serviços de estatística"

[Nota: sala das máquinas tabeladoras]

"No segundo pavimento estão os gabinetes da direção, a secretaria geral e várias repartições"
[Nota: gabinete do diretor geral do INE à época, Eng.º Tovar de Lemos]

"(...) formando ambiente de elegância e de comodidade indispensável onde logo ressalta a ordem e a disciplina que caracterizam este modelar organismo"
[Nota: sala das máquinas perfuradoras]

"Todas as repartições e gabinetes estão mobilados com gosto moderno"
[Nota: sala de trabalho do Comércio Interno]

"Toda a vida do país e do seu vasto Império Colonial pode ser, de momento compulsada no Palácio da Estatística que é um modelo de disciplina de trabalho em que se encontram expostos os mapas estatísticos de todas as nossas atividades que qualquer pessoa leiga que seja entende sem esforço"
[Nota: antiga sala de leitura da biblioteca]

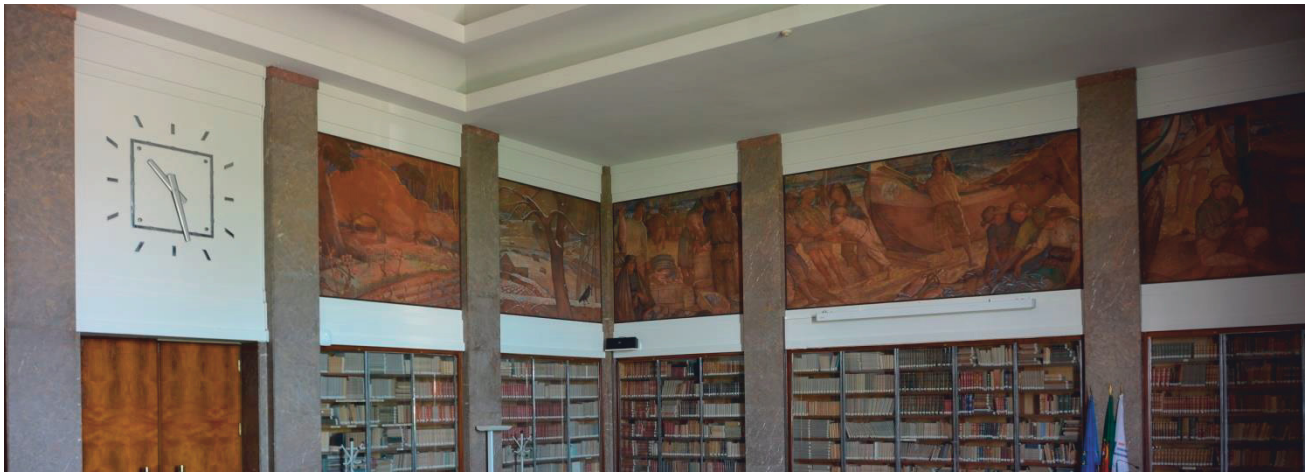
 (**)

"Os ficheiros são completíssimos e dos mais perfeitos que há em organismos deste género"
[Nota: uma das salas de trabalho dos serviços de estatística]

"As obras de arte não foram esquecidas como complemento ornamental do grande e majestoso edifício e devem-se a Abel Manta que desenhou um grande vitral executado por Ricardo Leone, simbolizando as fases da vida"

"Martinho da Fonseca fez a decoração do Salão Nobre onde se encontram estantes carregadas de livros artisticamente encadernados"
[Nota: nesta imagem do Salão Nobre, de 1942, observa-se um painel iniciado pelo pintor Martinho da Fonseca; posteriormente os trabalhos deste artista são retirados e deslocalizados; só em 1947 é que o pintor Henrique Franco executará os atuais dez painéis do Salão Nobre]

"Numa outra sala, destinada a reuniões das comissões técnicas, veem-se belos retratos do chefe de Estado, Presidente do Conselho, Dr. Armindo Monteiro – que foi diretor do Instituto Nacional de Estatística e do professor António Vilaça, primeiro diretor deste estabelecimento, todos devido ao pincel de Eduardo Malta"



[Nota: os dez painéis existentes no Salão Nobre, da autoria do pintor Henrique Franco, são alusivos às quatro estações do ano; há dois trípticos, um alegórico às atividades relacionadas com o mar e outro com a terra]

[Nota: no âmbito da arte integrada no edifício são de referir os dois baixos-relevos que encimam os janelões da fachada do edifício, da autoria do escultor Leopoldo de Almeida, um alusivo à Agricultura e Demografia e outro ao Comércio e Indústria]

"Todas as dependências do Edifício de Estatística revelam-se, a par do bom gosto que vai da sobriedade das suas linhas ao mobiliário, uma ordem e disciplina que poucas vezes se encontra em grau tão elevado, que por si distinguem os dirigentes deste Instituto"

[Nota: de referir alguns detalhes estéticos dos complementos de arquitetura associados à *art deco*, tais como as linhas direitas e geométricas dos elementos decorativos da fachada e de outros componentes no interior do edifício]

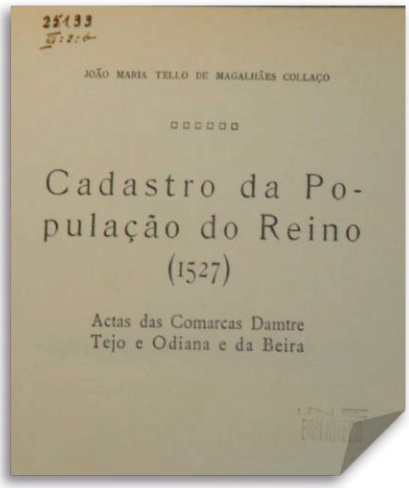
Guardar o tempo: o património bibliográfico

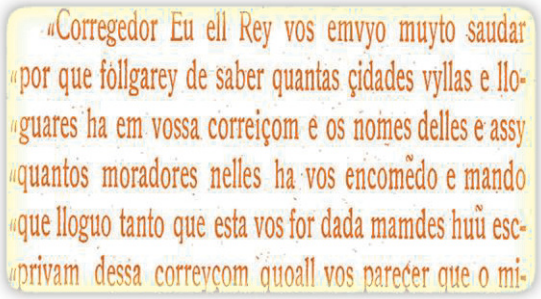
O acervo documental do Instituto Nacional de Estatística é rico e variado. Conta atualmente com cerca de 54 mil volumes distribuídos entre publicações estatísticas, monografias e revistas técnicas e científicas.
Enquanto biblioteca especializada em estatística (publicações de dados e documentação de suporte teórico), o INE integrou o legado documental dos organismos responsáveis pela estatística oficial que precederam o Instituto. Ao longo dos anos o seu património documental foi enriquecido com doações de outras instituições e até de particulares.
O INE possui alguns documentos setecentistas e oitocentistas de grande interesse documental, porém o acervo do séc. XIX é já bastante alargado com obras de cariz metodológico, sobre agricultura, população, fiscalidade, etc.
Na transição do séc. XIX-XX eram poucos os títulos ativos mas já havia alguma regularidade na sua publicação, como por exemplo: censos, anuário estatístico, contribuições e impostos, comércio internacional, demografia e pescas.
O século XX é o século da expansão da informação estatística publicada. Os temas estatísticos ampliam-se com vários títulos setoriais, a periodicidade regulariza-se e, em alguns casos, passa a ser de carácter infra anual; é neste período que os territórios ultramarinos passam a produzir publicações de modo consistente e regular.
Entramos no século XXI com uma grande abrangência temática e damos lugar ao formato eletrónico (disquetes, CD-ROM), até entrarmos na era da informação maioritariamente digital, customizada e permanentemente atualizada no Portal do INE.
Apresentamos, pois, o pequeno conjunto de obras de que muito nos orgulhamos de ser guardiões. Dado o seu valor documental, estas obras só estão disponíveis para consulta sob pedido, sendo os principais consulentes investigadores e académicos.

Salão Nobre do INE

Aquando da inauguração do edifício (c. 1935) o espaço era designado "salão da biblioteca" porque dispunha de estantes metálicas cravadas na parede onde, desde sempre, se guardou uma parte do património documental do INE; todavia, dada a dimensão do mesmo, o INE dispõe de um amplo espaço de arquivo onde está albergado o sempre crescente acervo

1527: Cadastro da População do Reino: Actas das Comarcas Damtre Tejo e Odiana e da Beira



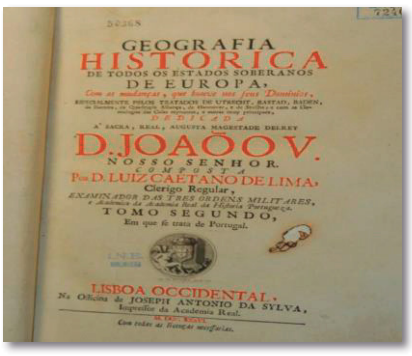


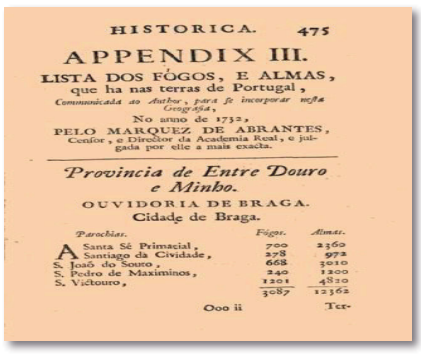
Em termos de data de referência dos dados (1527) é o documento mais antigo e é considerado uma "contagem" ou "numeramento" da população. O que o INE possui é uma transcrição realizada por um académico de Coimbra, Prof. Magalhães Colaço, que a publicou pela primeira vez em 1929, em edição de autor.

O original desapareceu, foi roubado da Torre do Tombo, e depois vendido a um oficial inglês e encontra-se no British Museum, em Londres. Em Coimbra existe uma cópia e foi a partir dela que o Prof. Magalhães Colaço, professor de estatística naquela universidade, fez a transcrição que publicou em 1929. Posteriormente transferiu-se para a Universidade de Lisboa e, em 1934, o documento foi publicado na Revista da Faculdade de Direito de Lisboa, numa 2ª edição.

O INE possui um exemplar de cada uma das referidas edições. É um documento de consulta difícil pela linguagem arcaica e pela nomenclatura territorial ser tão distinta da atual, o que dificulta comparações. As variáveis são: "moradores" e "fogos", embora também apareça a designação de "vizinhos". A outra dificuldade é que não cobre a totalidade do território, mas apenas as "comarcas Damtre Tejo e Odiana e da Beira" ou seja, do Douro para baixo.

1732: Lista dos fogos, e almas, que há nas terras de Portugal





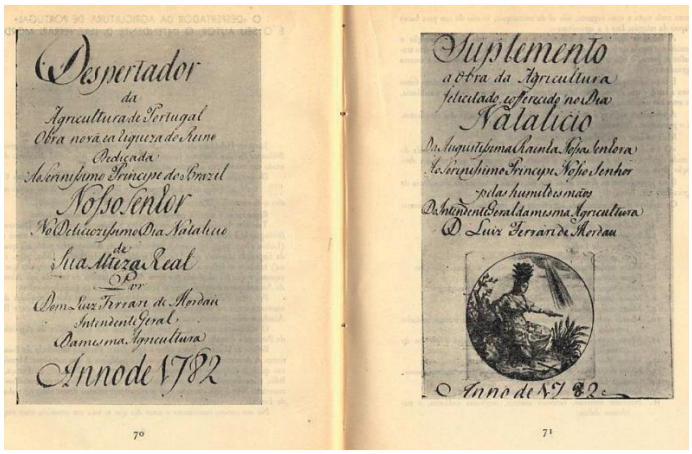
A "Lista dos fogos e almas que há nas terras de Portugal" é um documento datado de 1736 e foi incluído no "Apêndice III" da obra «Geografia histórica de todos os Estados soberanos da Europa…, tomo segundo, em que se trata de Portugal». É considerado um "numeramento" e só tem duas variáveis.

A «Geografia Histórica…» foi realizada por um clérigo, D. Luís Caetano de Lima, examinador das três ordens militares e membro da Academia Real de História Portuguesa; foi dedicada a D. João V.

A "Lista dos fogos e almas que há nas terras de Portugal" tem sido alvo de alguma reserva por parte dos investigadores por nunca ter sido encontrado o original e não haver forma de validar o documento que chegou aos nossos dias.

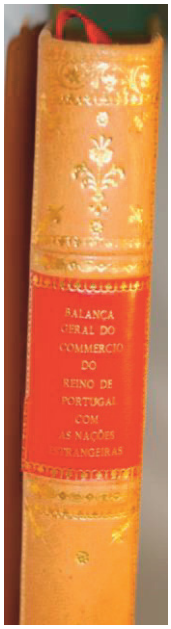
O "numeramento" é de difícil consulta por ter como referência territorial uma nomenclatura distinta da atual. Contem apenas duas variáveis, como é indicado no título: "fogos" e almas".

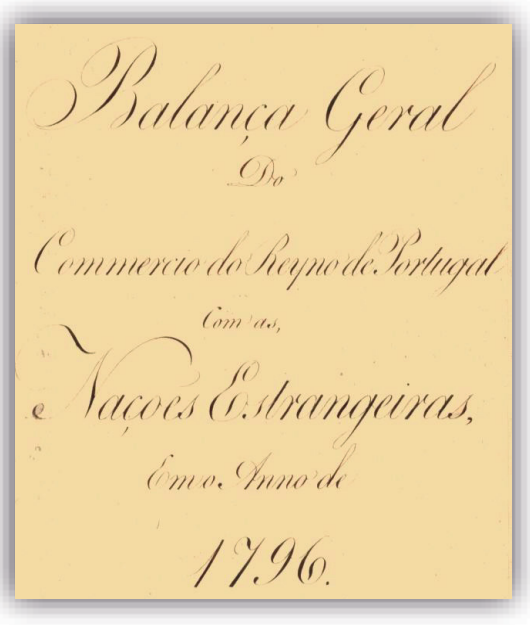
1782: Despertador da agricultura de Portugal: obra nova e a riqueza do Reino: dedicada ao sereníssimo príncipe do Brasil, nosso senhor, no deliciosíssimo dia natalício de sua alteza real



A obra é da autoria de Luís Ferrari Mordau, cidadão espanhol, que veio viver e trabalhar para Portugal. Em 1765, durante o reinado de D. José I, foi nomeado Intendente Geral da Agricultura, mas o seu trabalho foi mormente desenvolvido durante o reinado de D. Maria I. A obra original é de 1782.
Em 1950, o Prof. Moses Bensabat Amzalak reproduziu o trabalho de Mordau num estudo publicado na «Revista do Centro de Estudos Económicos», editado pelo INE, em 1950. No acervo do INE existe apenas o documento reproduzido no estudo do Prof. Amzalak.

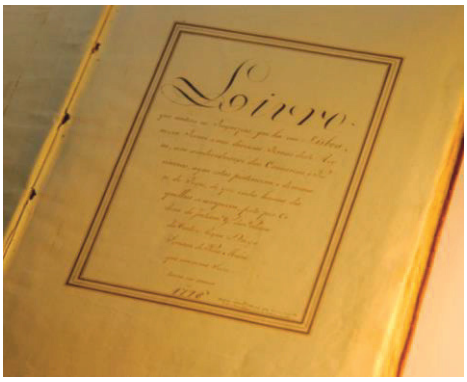
1796: Balança geral do comércio do reino de Portugal com os seus domínios e nações estrangeiras





Esta série documental, da responsabilidade daquela que pode ser considerada a primeira instituição oficial de estatística de Portugal, ainda que setorial (Superintendência Geral dos Contrabandos e Descaminhos dos Reais Direitos), apresenta rigor, sistematização e continuidade no tempo.
É o periódico mais antigo em Portugal e não deve haver muitos iguais no mundo; o exemplar mais antigo em posse do INE refere-se às estatísticas do comércio externo de 1796. O seriado foi editado até 1831 e ainda hoje existe com a designação de "estatísticas do comércio internacional".
Na coleção existente no INE, os livros das "balanças do comércio" são manuscritos e a apresentação dos dados está dividida em dois volumes: um para os Domínios (ultramarinos) e outro para as Nações estrangeiras; tem uma nomenclatura de "bens, quantidades e valor", em reis.
É um documento muito pretendido pelos investigadores.

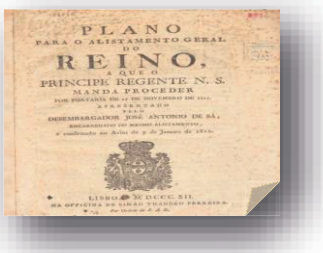
1798: Livro que contém as freguesias que há em Lisboa, no seu termo, e nas diversas terras deste reino, com a individuação das Comarcas e Províncias, a que estas pertencem, e do número de fogos, de que cada uma daquelas se compõem, feito por ordem do Intendente Geral da Polícia da Corte, e Reino, Diogo Inácio de Pina Manique, em sua Secretaria no ano de 1798



Este "numeramento" foi realizado durante o reinado de D. Maria I, pelo Intendente Geral da Polícia da Corte e Reino, Diogo Inácio de Pina Manique.
Vulgarmente conhecido por "censo de Pina Manique", o seu conteúdo resulta de um levantamento de fogos com o objetivo claramente militar de recompor o exército com novos recrutas. O texto original apresenta duas variáveis "fogos e "recrutas".
A obra é manuscrita e constituída por 43 páginas e/ou fólios, a que acresce uma recapitulação geral que engloba todas as províncias.
Em 1970, através do Centro Cultural Português, da Fundação Calouste Gulbenkian, em Paris, o historiador Joaquim Veríssimo Serrão publicou um estudo sobre este "numeramento" intitulado «A população de Portugal em 1798: o censo de Pina Manique», podendo ser encontrados elementos adicionais para a compreensão da obra.
Na atualidade, os especialistas arranjaram um multiplicador que permitiu estimar a população à época (3,5 - 5 pessoas/fogo) e chegaram, assim, a um valor estimado da população em 1798.

1801: Instruções gerais para se formar o cadastro, ou o mapa aritmético-político, do reino feitas por ordem de sua Alteza real o príncipe regente nosso senhor

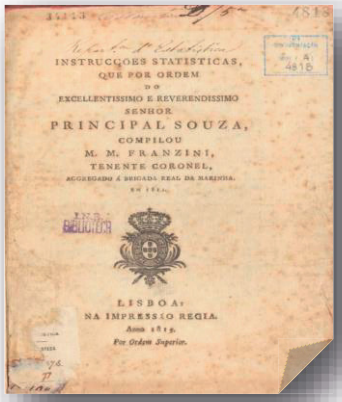
1812: Plano para o alistamento geral do Reino, a que o Príncipe regente N. S. manda proceder por portaria de 15 de novembro de 1811



Em 1801 já havia preocupações metodológicas com a estatística. José António de Sá, que foi desembargador (juiz) e superintendente geral da décima da Corte do Reino, é autor de duas importantes obras de carácter metodológico.
A primeira surge precisamente em 1801 com o título de «Instruções gerais para se formar o cadastro, ou o mapa aritmético-político (…)»; a expressão "aritmética-política" é, em linguagem da época, aquilo que atualmente se designa por "estatística".

A segunda obra surge em 1812, intitulada «Plano para o alistamento geral do Reino (…)»; é já um documento de ordem prática, tratando-se de um "plano de execução". Esta obra completa a anterior e propõe "tabelas de recolha de dados".
Em 1945 o INE publicou estas duas obras integradas num coleção denominada «Subsídios para a história da estatística em Portugal: volume I» e com o título genérico de "Cadastro do reino: 1801-1812".

1814: Instruções estatísticas, que por ordem do Exmo. e Revmo. Sr. Principal Sousa, compilou M. M. Franzini, tenente-coronel, agregado à brigada Real da Marinha, em 1814


INSTRUCÇÕES STATISTICAS,
QUE POR ORDEM
DO
EXCELLENTISSIMO E REVERENDISSIMO
SENHOR
PRINCIPAL SOUZA,
COMPILOU
M. M. FRANZINI,
TENENTE CORONEL,
AGREGADO Á BRIGADA REAL DA MARINHA.
EM 1814.
LISBOA:
NA IMPRESSÃO REGIA.
Anno 1815.
Por Ordem Superior.

Esta obra foi publicada em 1815. O responsável pela sua compilação foi Marino Miguel Franzini, muito conhecido no meio da estatística do séc. XIX. Tinha ascendência italiana (o pai era catedrático em Coimbra) e foi tenente-coronel do exército e, mais tarde, diretor do Arquivo Militar; fez parte da Comissão Central de Estatística (c. 1840-1850).
As "instruções estatísticas" foram ordenadas pelo Principal de Sousa ("Principal" é um termo eclesiástico para designar o superior de uma comunidade religiosa). O Principal de Sousa, de seu nome José António de Menezes de Sousa Coutinho, foi membro do conselho de regência que ficou a governar Portugal aquando do exílio da família real no Brasil (D. Maria I e o príncipe regente, futuro D. João VI).
Anos depois de publicar as «Instruções estatísticas», Franzini publicou uma mini nomenclatura de profissões e atividades da população, uma obra de caráter ensaístico.

FONTES

"O Palácio da Estatística"

- «Panorama, Revista Portuguesa de Arte e Turismo», nº 11, 1942, Hemeroteca Municipal de Lisboa, http://hemerotecadigital.cm-lisboa.pt/
- «INE 80 anos: um outro olhar», Instituto Nacional de Estatística, 2015.
- «O Instituto Nacional de Estatística», Instituto Nacional de Estatística, 1936.

Guardar o tempo: o património bibliográfico

- «Guardar o tempo: mostra histórica de documentos», Instituto Nacional de Estatística, 2013.

IMAGENS

"O Palácio da Estatística"

- «Panorama, Revista Portuguesa de Arte e Turismo», nº 11, 1942. (assinaladas com **)
- Banco de Imagens do INE.

Guardar o tempo: o património bibliográfico

- Banco de Imagens do INE.

Ciência Estatística

• Artigos em Revistas

Brites, N. M. e Braumann, C. A. (2020). Stochastic differential equations harvesting policies: Allee effects, logistic-like growth and profit optimization. *Appl. Stochastic Models Bus. Ind.*. 2020;1–11.

Papança, F (2019). Anastácio da Cunha, O Paço da Bemposta, capela e espólio da Biblioteca da Academia Militar; livros de Matemática e de Estatística utilizados na formação de oficiais do Exército no período de 1640-1926.
Revista Militar, II Século -71º Volume - nº 11. Nº 2614. p. 1155-1191

• Capítulos de Livros

Título: *Harvesting Policies with Stepwise Effort and Logistic Growth in a Random Environment*
Autores: Brites, N. M. e Braumann, C. A.
Livro: *Current Trends in Dynamical Systems in Biology and Natural Sciences*
Editores: Aguiar M., Braumann C., Kooi B., Pugliese A., Stollenwerk N., Venturino E.
Ano: 2020. Edição: SEMA SIMAI Springer Series, vol 21; p. 95-110.
ISBN: 978-3-030-41119-0; 978-3-030-41120-6 (eBook)

Título: *0 ensino, a aprendizagem e a utilização da Estatística: da teoria à prática*
Autores: Áurea Sousa
Livro: *Temas cruzados. Pensamentos interligados*
Editores: A. Mendes, Á. Sousa, H. Melo, J. Nunes, M. C. Martins, O. Silva & P. Medeiros.
Ano: 2019. Edição: Universidade dos Açores, pp. 54-71.
ISBN: 978-989-33-0051-0

• Teses de Mestrado

Título: Modelo de previsão de vendas para otimização do reaprovisionamento do produto às lojas
Autora: Vanessa Bandeira Penso, *vanessabpenso@gmail.com*
Orientadoras: Maria do Carmo Miranda Guedes e Margarida Maria Araújo Brito

• Livros

Título: *ATAS DO XXIII CONGRESSO da Sociedade Portuguesa de Estatística*
Editores: Maria de Fátima Salgueiro, Paula Vicente, Teresa Calapez, Catarina Marques e Maria Eduarda Silva
Ano: 2020. Edições SPE.
ISBN: 978-972-8890-46-9

Título: *Current Trends in Dynamical Systems in Biology and Natural Sciences*
Editores: Aguiar M., Braumann C., Kooi B., Pugliese A., Stollenwerk N., Venturino E.
Ano: 2020. Edição: SEMA SIMAI Springer Series, vol 21.
ISBN: 978-3-030-41119-0; 978-3-030-41120-6 (eBook)
Nota: Contém Prefácio dos Editores

• Doutoramento

Título: Latent Class Models in the Evaluation of Biomedical Diagnostic Tests and Internet Traffic Anomaly Detection

Autora: Ana Patrícia Subtil da Graça Freitas Garcia, *anasubtil@tecnico.ulisboa.pt*
Orientadores: Maria do Rosário Oliveira e António Pacheco

A minha tese tem como fio condutor os modelos de classes latentes (MCL), modelos estatísticos que relacionam variáveis observáveis com variáveis subjacentes não directamente observáveis, ditas latentes. A tese foca dois problemas práticos, a avaliação de desempenho de testes de diagnóstico biomédico e a detecção de ataques de redirecionamento de tráfego na Internet, que permitiram explorar os MCL sob diferentes perspectivas.

A tese aborda o problema da estimação da sensibilidade e especificidade, medidas convencionais de desempenho de um teste de diagnóstico dicotómico, na ausência de uma referência perfeita. Nestas circunstâncias, a sensibilidade e a especificidade podem ser estimadas recorrendo a métodos baseados em testes de diagnóstico imperfeitos: comparação com um teste imperfeito, comparação com uma combinação de testes imperfeitos, análise de discrepâncias e MCL. Foram derivadas expressões analíticas para os enviesamentos da sensibilidade e da especificidade e explorado o impacto nesses enviesamentos de factores como a prevalência da condição e a dependência condicional entre os testes. Este estudo revelou que os métodos baseados na comparação com referências imperfeitas são generalizações do caso particular da comparação com um único teste imperfeito. A estratégia de estudar de forma integrada diferentes métodos de estimação da sensibilidade e especificidade conduziu a resultados comparáveis e consistentes. A natureza teórica da abordagem conseguiu superar as limitações resultantes do uso de informação parcial inerente aos estudos por amostragem. Não foi possível identificar um método capaz de minimizar os enviesamentos em todas as circunstâncias, pelo que não conseguimos fornecer uma recomendação única. Consequentemente, foi desenvolvida uma ferramenta interactiva disponibilizada na Internet, que permite a visualização da magnitude e direcção dos enviesamentos, em função de parâmetros de entrada indicados pelo utilizador. Médicos, biólogos ou outros investigadores podem consultar esta ferramenta para obterem, de modo simples e interactivo, informação rigorosa conducente a escolhas fundamentadas de metodologias de avaliação de testes de diagnóstico, na ausência de uma referência perfeita.

A tese abordou também o problema prático da detecção de ataques de redirecionamento na Internet, em que um um agente malicioso desvia secretamente o tráfego entre dois servidores. Foram propostas metodologias de detecção destes ataques com base nos tempos de ida e volta (RTT, *round-trip-times*) medidos periodicamente entre um conjunto de dispositivos de origem de medição, distribuídos geograficamente, e um servidor alvo, potencial vítima dos ataques. A abordagem ensaiada recorre a algoritmos de aprendizagem supervisionada (kNN, C5.0, random forest e AdaBoost) para, atendendo aos RTT medidos, classificar como regular ou anómalo o tráfego entre cada um dos múltiplos dispositivos de origem e o alvo monitorizado. As classificações correspondentes aos diferentes dispositivos de origem são posteriormente combinadas usando MCL e modelos de Markov escondidos (HMM, *hidden Markov models*), para decidir se o servidor sob vigilância está ou não a sofrer um ataque. A metodologia inovadora proposta foi aplicada a diversos conjuntos de dados e conduziu a níveis de desempenho elevados na detecção de ataques de redirecionamento. Estes resultados sugerem que a aplicação de MCL e HMM na combinação de classificadores é uma estratégia promissora.

Ana Subtil

Edições SPE - Mini Cursos

Título: *Uma introdução à Meta-Análise*
Autora: Maria de Fátima Brilhante
Ano: 2017.

Título: *Estatística Bayesiana Computacional – uma introdução*
Autores: M. Antónia Amaral Turkman e Carlos Daniel Paulino
Ano: 2015.

Título: *Análise de Valores Extremos: Uma Introdução*
Autoras: M. Ivette Gomes, M. Isabel Fraga Alves e Claudia Neves
Ano: 2013.

Título: *Modelos com Equações Estruturais*
Autora: Maria de Fátima Salgueiro
Ano: 2012.

Título: *Análise de Dados Longitudinais*
Autoras: Maria Salomé Cabral e Maria Helena Gonçalves
Ano: 2011

Título: *Uma Introdução à Estimação Não-Paramétrica da Densidade*
Autor: Carlos Tenreiro
Ano: 2010

Título: *Análise de Sobrevivência*
Autoras: Cristina Rocha e Ana Luísa Papoila
Ano: 2009

Título: *Análise de Dados Espaciais*
Autoras: M. Lucília de Carvalho e Isabel C. Natário
Ano: 2008

Título: *Introdução aos Métodos Estatísticos Robustos*
Autores: Ana M. Pires e João A. Branco
Ano: 2007

Título: *Outliers em Dados Estatísticos*
Autor: Fernando Rosado
Ano: 2006

Título: *Introdução às Equações Diferenciais Estocásticas e Aplicações*
Autor: Carlos Braumann
Ano: 2005

Título: *Uma Introdução à Análise de Clusters*
Autor: João A. Branco
Ano: 2004

Título: *Séries Temporais – Modelações lineares e não lineares*
Autoras: Esmeralda Gonçalves e Nazaré Mendes Lopes
Ano: 2003 (2ª Edição em 2008)

Título: *Modelos Heterocedásticos. Aplicações com o software Eviews*
Autor: Daniel Muller
Ano: 2002

Título: *Inferência sobre Localização e Escala*
Autores: Fátima Brilhante, Dinis Pestana, José Rocha e Sílvio Velosa
Ano: 2001

Título: *Modelos Lineares Generalizados – da teoria à prática*
Autores: M. Antónia Amaral Turkman e Giovani Silva
Ano: 2000

Título: *Controlo Estatístico de Qualidade*
Autoras: M. Ivette Gomes e M. Isabel Barão
Ano: 1999

Título: *Tópicos de Sondagens*
Autor: Paulo Gomes
Ano: 1998

Prémio SPE 2020

Prémio SPE 2020

Pretendendo estimular a atividade de estudo e investigação científica em Probabilidades e Estatística, a Sociedade Portuguesa de Estatística institui em 2020 o Prémio SPE, regido pelo seguinte regulamento.

Está aberto até 15 de Setembro de 2020 o concurso para atribuição do Prémio SPE 2020, doravante referido como prémio. O prémio é constituído por uma quantia de 1000 euros.

Ao prémio podem concorrer trabalhos originais sobre temas de Probabilidades e Estatística, desde que não tenham sido objeto de qualquer prémio atribuído por outra instituição. O trabalho deverá ser apresentado em português ou em inglês e não poderá exceder 25 páginas A4.

Podem candidatar-se ao prémio jovens investigadores sócios da SPE que não completem 35 anos de idade até 31 de Dezembro de 2020 ou que tenham obtido doutoramento há menos de 5 anos e que não tenham recebido o prémio nas quatro edições anteriores.

A decisão de admissibilidade e a apreciação dos trabalhos submetidos a concurso é da competência de um Júri, cuja constituição é da responsabilidade da Direção da SPE.

Os critérios de seleção pautar-se-ão pela exigência e precisão nos vários aspetos que o Júri considerar pertinentes, nomeadamente: i) qualidade e clareza do texto; ii) inovação e rigor científico; iii) contribuição para o desenvolvimento da área de Probabilidades e Estatística nos planos teórico, metodológico e/ou aplicado.

O Júri é soberano nas suas decisões, não havendo lugar a recurso.

O Júri reserva-se o direito de não atribuir o Prémio SPE 2020.

As candidaturas ao prémio, dirigidas à Presidente da SPE, são constituídas pelos trabalhos concorrentes e pelo *curriculum vitae* dos autores. Devem ser enviadas por correio eletrónico para spe@spestatistica.pt

A entrega formal do Prémio SPE 2020, com apresentação do trabalho galardoado, terá lugar durante o XXV Congresso da Sociedade Portuguesa de Estatística.

O MUNDO DA ESTATÍSTICA

ORGANIZAÇÃO PARTICIPANTE

Índice

Editorial 1

Mensagem da Presidente 3

Notícias 4

Enigmística 7

INE - 85 anos de estatísticas a servir o país

Mensagem do Presidente do Instituto Nacional de Estatística

Francisco Lima *8*

Inovação nas Estatísticas Oficiais - Principais desafios do INE

Instituto Nacional de Estatística *10*

Modelos de regressão para dados de contagem: uma aplicação ao Inquérito ao Transporte Rodoviário de Mercadorias

Inês Rodrigues *14*

Short-Term Regional Demographic Forecasts with Time Series Methods and Machine Leaming Algorithms

Jorge M Bravo e Edviges Coelho *20*

Support Vector Machine para Imputação e Edição de Valores - O caso das Declarações Mensais de Remuneração das Empresas

Filipe Santos e Pedro Campos *30*

O papel da intermediação do setor financeiro - quando o todo é menor do que a soma das partes

Filipa Lima, Sónia Mota e Ângela Coelho *36*

O INE como (pro)motor da Literacia Estatística

Francisco Correia *42*

INE: preservar a memória institucional

Paula Marques *46*

Ciência Estatística

Livros e Capitulos de Livros 57

Teses de Doutoramento 58

Edições SPE - Mini Cursos 59

"Prémio SPE" 60

Este boletim tem o apoio de:

FCT Fundação para a Ciência e a Tecnologia

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR