Universidade de Lisboa

Instituto de Geografia e Ordenamento do Território

Padrões Agroecológicos Históricos em Portugal Continental: Uma Abordagem com Inteligência Artificial e SIG

Diogo Filipe Ferreira de Carvalho

Mestrado em Sistemas de Informação Geográfica e Modelação Territorial Aplicados ao Ordenamento

Dissertação de Mestrado Orientada pela Doutora Cláudia Morais Viana e pelo Doutor Fernando Jorge Pedro da Silva Pinto da Rocha

2024

Universidade de Lisboa

Instituto de Geografia e Ordenamento do Território

Padrões Agroecológicos Históricos em Portugal Continental: Uma Abordagem com Inteligência Artificial e SIG

Diogo Filipe Ferreira de Carvalho

Mestrado em Sistemas de Informação Geográfica e Modelação Territorial Aplicados ao Ordenamento

Dissertação de Mestrado Orientada pela Doutora Cláudia Morais Viana e pelo Doutor Fernando Jorge Pedro da Silva Pinto da Rocha

Júri:

Presidente: Doutor Nuno Manuel Sessarego Marques da Costa, Professor Associado do Instituto de Geografia e Ordenamento do Território da Universidade de Lisboa

Vogais:

· Doutora Raquel Margarida Viana Faria de Deus, Investigadora do CICS.NOVA - Centro Interdisciplinar de Ciências Sociais e do e-GEO - Centro de Estudos de Geografia e Planeamento Regional da Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa

· Doutor António Manuel Teixeira Monteiro, Investigador Auxiliar do Instituto de Geografia e Ordenamento do Território da Universidade de Lisboa

· Doutora Cláudia Morais Viana, Investigadora Júnior do Instituto de Geografia e Ordenamento do Território da Universidade de Lisboa

2024

Declaração de Autoria

Eu Diogo Filipe Ferreira de Carvalho, declaro que a presente dissertação de mestrado intitulada *Padrões Agroecológicos Históricos em Portugal Continental: Uma Abordagem com Inteligência Artificial e SIG*, é o resultado da minha investigação pessoal e independente. O conteúdo é original e todas as fontes consultadas estão devidamente mencionadas na bibliografia ou outras listagens de fontes documentais, tal como todas as citações diretas ou indiretas têm devida indicação ao longo do trabalho segundo as normas académicas.

Esta dissertação foi elaborada no âmbito do Projeto AgroEcoDecipher - Análise geográfica das tendências agroecológicas através de fontes históricas e inteligência artificial, financiado pela Fundação para a Ciência e Tecnologia (2022.09372.PTDC).

Agradecimentos

Não seria justo começar os meus agradecimentos sem fazer menção primeiramente aos meus pais por todo o carinho, apoio e amor que me deram ao longo desta etapa. Sem os seus sacrifícios não teria a oportunidade de terminar este trabalho e por isso o culminar desta fase académica é também um pouco deles.

Abaixo, expresso meu agradecimento a pessoas especiais, que contribuíram de forma significativa neste último ano de trabalho.

À Bi agradeço pelo incentivo constante, apoio e paciência para ouvir os meus desabafos e frustrações.

Aos meus amigos: ao André, que há mais de 10 anos que me incentiva e me apoia nas minhas conquistas; ao Pedro Tomás pelo constante apoio ao longo desta jornada. Agradeço também ao Diogo, ao Pedro, à Beatriz, à Mariana, à Bárbara e à Raquel pela entreajuda, preocupação e interesse (embora alguns não percebessem exatamente o que eu estava a fazer).

Por fim, quero expressar meu profundo agradecimento aos meus orientadores, não apenas pelas qualidades profissionais, amplamente reconhecidas por todos, mas também pelas qualidades pessoais, que nem todos têm a oportunidade de conhecer. À Doutora Cláudia Viana, deixo minha gratidão por todos os ensinamentos, pela presença constante ao longo de todo o trabalho, pela compreensão diante dos erros e pelo incentivo à comunicação aberta e simples. Nunca faltaram palavras de apoio e motivação, disponibilizando-se sempre para ouvir novas ideias e oferecer sugestões independentemente da hora do dia. Gostaria de destacar um momento em particular: durante uma conferência internacional, quando estava nervoso por falar diante de tantas pessoas, a Doutora Cláudia foi a primeira a perceber meu nervosismo e a dar-me forças para enfrentar aquele desafio. Este trabalho jamais estaria concluído sem as oportunidades que me deu.

Ao professor Doutor Jorge Rocha quero agradecer por todas as ideias e sugestões dadas para o trabalho, arranjando sempre tempo para me auxiliar independentemente dos numerosos projetos onde estava inserido. Foi um gosto enorme trabalhar de perto com alguém cujo trabalho admiro desde que tive o meu primeiro contacto com os Sistemas de Informação Geográfica no IGOT.

Resumo

Esta dissertação explora como a Inteligência Artificial (IA) e os Sistemas de Informação Geográfica (SIG) podem contribuir para a análise de padrões agroecológicos históricos em Portugal Continental na década de 1950. De forma inovadora, combina Reconhecimento Ótico de Caracteres (OCR) e modelos de linguagem avançados, como o ChatGPT, para melhorar a extração e análise de dados históricos. Utilizando técnicas de NLP, text mining e aprendizagem automática, analisa 163 inquéritos agrícolas para compreender especificidades regionais e criar mapas agrícolas detalhados.

A estrutura da dissertação abrange cinco capítulos. O Capítulo I apresenta o enquadramento teórico, destacando a relevância do estudo no contexto das alterações climáticas e da necessidade de práticas agrícolas mais sustentáveis. No Capítulo II, examina-se o papel da IA e dos SIG no apoio às humanidades digitais, com foco nas aplicações de OCR, Spatial Data Mining e ChatGPT para a análise de textos históricos. O Capítulo III detalha os dados e a metodologia, incluindo a digitalização, tratamento cartográfico e análise textual. O Capítulo IV apresenta os resultados, destacando a produção de mapas agrícolas e a identificação de padrões agroecológicos regionais. Por fim, no Capítulo V, conclui-se com as implicações desta abordagem interdisciplinar para a agricultura sustentável.

A principal inovação reside na aplicação pioneira, em Portugal, da integração de OCR e ChatGPT para análise de fontes históricas. Esta combinação permite uma extração e interpretação de dados mais eficiente e precisa, contribuindo para o avanço da investigação histórica e oferecendo novas perspetivas para o desenvolvimento agrícola. Além de reforçar o uso de tecnologias modernas em contextos históricos, esta abordagem tem potencial impacto internacional, promovendo novas formas de estudar e interpretar o passado agrícola para orientar práticas futuras.

Palavras-Chave: K-means, Reconhecimento Ótico de Caracteres, Análise espacial, Humanidades digitais, Informação histórica

Abstract

This dissertation explores how Artificial Intelligence (AI) and Geographic Information Systems (GIS) can contribute to analyzing historical agroecological patterns in mainland Portugal during the 1950s. Innovatively, it combines Optical Character Recognition (OCR) and advanced language models, such as ChatGPT, to enhance the extraction and analysis of historical data. Using techniques such as Natural Language Processing (NLP), text mining, and machine learning, the study analyzes 163 agricultural surveys to understand regional specificities and create detailed agricultural maps.

The dissertation is structured into five chapters. Chapter I introduces the theoretical framework, emphasizing the study's relevance in the context of climate change and the need for more sustainable agricultural practices. Chapter II examines the role of AI and GIS in supporting digital humanities, focusing on the applications of OCR, Spatial Data Mining, and ChatGPT for analyzing historical texts. Chapter III outlines the data and methodology, including digitization, cartographic processing, and textual analysis. Chapter IV presents the results, highlighting the production of agricultural maps and the identification of regional agroecological patterns. Finally, Chapter V concludes with the implications of this interdisciplinary approach for sustainable agriculture.

The primary innovation lies in the pioneering application of OCR and ChatGPT integration for historical data analysis in Portugal. This combination enables more efficient and accurate data extraction and interpretation, advancing historical research and offering new perspectives for agricultural development. Beyond enhancing the use of modern technologies in historical contexts, this approach has potential international significance, promoting novel ways of studying and interpreting agricultural history to guide future practices.

Keywords: K-means, Optical Character Recognition, Spatial Analysis, Digital Humanities, Historical Information

Índice

Índice de Figuras

Índice de Quadros

´

Capítulo 1 - Introdução

1.1. Enquadramento do tema

As consequências das alterações climáticas, como secas, inundações e incêndios florestais, exercem um impacto substancial sobre a agricultura e a segurança alimentar (Yadav, 2019). Estes desafios obrigam a uma vigilância atenta, uma vez que, por exemplo, os padrões de precipitação estão sujeitos a alterações tanto no espaço como no tempo (Chou *et al.*, 2013), com tendência prevista para persistir no futuro próximo (Chou & Lan, 2012). Os países altamente desenvolvidos estão particularmente vulneráveis a estas mudanças, uma vez que as suas práticas agrícolas em larga escala dependem diretamente da precipitação para garantir a eficiência da produção. Por outro lado, regiões do continente africano enfrentam desafios adicionais devido às suas condições intrínsecas, incluindo solos inadequados para a agricultura e um clima extremamente seco em grande parte do continente (Arndt *et al.*, 2012). Desta forma, as alterações significativas no clima global, aliadas, por exemplo, à diminuição da disponibilidade e qualidade da água, representam desafios prementes no cenário mundial contemporâneo. Esta realidade impõe uma necessidade urgente de redefinição dos paradigmas de gestão e desenvolvimento de recursos essenciais, sobretudo devido às projeções de crescimento populacional exponencial (Carvalho, 2006). Tais desafios abrangem uma nova perspetiva nas atividades humanas, entre as quais se destaca a agricultura.

Historicamente, a agricultura tem sido um pilar fundamental da economia global, enfrentando uma pressão constante para elevar a sua produtividade e eficiência, em resposta ao crescente aumento populacional (Ayed & Hanana, 2021). Desde a realização da Organização Mundial de Alimentação em 1996, têm sido realizados esforços para aumentar a produção e a segurança alimentar (FAO, 2017). Num dos 17 Objetivos de Desenvolvimento Sustentável, estabelecidos pelas Nações Unidas, está incluída a erradicação da fome. Contudo, as medidas implementadas têm-se revelado inadequadas devido, por exemplo, ao crescimento populacional e à aceleração da urbanização (Tomlinson, 2013). O desenvolvimento da sociedade moderna também transformou os padrões de consumo, resultando num aumento generalizado na ingestão calórica por pessoa, com dietas baseadas em proteína animal (Godfray *et al.*, 2010). Prevê-se o que a população continue a crescer (Lee, 2011), o que exige uma exploração agrícola ainda

mais intensiva para suprir a crescente necessidade de alimentos, prejudicando substancialmente a qualidade dos solos (Tóth *et al.*, 2018). Como tal, a adoção de práticas de agricultura sustentável torna-se imperativa.

Durante o século XX, a comunidade científica direcionou sua atenção para estes desafios emergentes, investigando as interações entre o ser humano e o ambiente com maior rigor. Atualmente, é amplamente aceite que a mudança de rumo requer uma cooperação global entre todas as nações a fim de desenvolver medidas de mitigação. A agricultura é uma prática com mais de 10.000 anos de história (Tauger, 2010), e, como tal, o controlo desta prática e a disponibilização de informações relacionadas à disponibilidade e produção de alimentos ao longo do tempo é crucial e poderá contribuir para desenvolver novas medidas de mitigação. Por exemplo, a existência de informações históricas possibilita uma melhor compreensão dos fenómenos naturais e sociais associados à agricultura no seu contexto espacial e temporal, permitindo a identificação de padrões que se desenvolveram ao longo do espaço e tempo, bem como contribuindo para modelos de possíveis tendências futuras (Knowles, 2005). Como tal, uma abordagem promissora para a elaboração de algumas medidas de mitigação e preparação para o futuro envolve a análise detalhada das dinâmicas históricas entre o ser humano e o ambiente. No contexto do estudo e identificação de possíveis tendências agrícolas futuras torna-se crucial considerar dados históricos, uma vez que as mesmas se relacionam diretamente com os fatores ambientais e naturais que se alteram ao longo do tempo e do espaço (Viana *et al.*,2021; Boivin & Crowther, 2021). Por outro lado, a reconstrução das tendências agroecológicas oferece uma visão mais completa do sistema agrícola do passado, enquanto revela, por exemplo, fatores ecológicos essenciais para garantir a segurança alimentar.

O crescente acesso a fontes históricas facilita a reconstrução de tendências agroecológicas, permitindo a análise da atividade humana e das mudanças nas práticas agrícolas face às atuais alterações globais. Esta abordagem oferece insights sobre o impacto da atividade agrícola no planeta e na segurança alimentar (Viana *et al.*, 2022). No entanto, existem vários desafios relacionados à utilização destas fontes, como o acesso e a qualidade das mesmas, que frequentemente estão disponíveis apenas em formato analógico e, devido à sua antiguidade, podem sofrer deterioração, comprometendo a integridade dos registos. Perante estas limitações, é essencial desenvolver metodologias eficazes para as superar. A utilização de tecnologia avançada, como a Inteligência

Artificial (IA), incluindo Reconhecimento Ótico de Caracteres (OCR), algoritmos de compreensão textual, modelos de Linguagem de Grande Escala (LLM), e Sistemas de Informação Geográfica (SIG), pode ser uma solução viável para contornar estes desafios. Estas tecnologias têm a capacidade de lidar com dados de diferentes estruturas e volumes, possibilitando a exploração e a combinação de diversos recursos de dados para modelar relações não lineares (Viana *et al.,* 2021). Em particular, os SIG desempenham um papel crucial neste contexto, pois auxiliam no tratamento de fontes históricas, na disponibilização dos dados e na condução de diversas análises relacionadas às mudanças espácio-temporais (Gatta *et al.*,2017; Knowles, 2008; Gregory & Southall, 2005). Outra vantagem na utilização destas tecnologias relaciona-se com a capacidade de modelar, com um grau de precisão relevante, variáveis associadas à produtividade agrícola (Ayed & Hanana, 2021). Também a IA e outras tecnologias digitais são potencialmente cruciais. Algoritmos de Aprendizagem Automática (*Machine learnin*g) e Aprendizagem Profunda (*Deep Learning*), conjugados com tecnologias avançadas de geolocalização, são exemplos de ferramentas essenciais no futuro da atividade agrícola como no estudo *Big Data Analysis for Sustainable Agriculture on a Geospatial Cloud Framework* (Delgado *et al.*, 2019). A emergência da Aprendizagem Automática, em sinergia com as tecnologias de *Big data*, proporciona uma oportunidade singular para quantificar, qualificar e encontrar novos caminhos na gestão dos recursos agrícolas (Liakos *et al.*, 2018). A implementação de soluções baseadas em IA no âmbito da agricultura tem demonstrado eficácia, contribuindo significativamente para o apoio à tomada de decisões e formulação de estratégias.

1.2. Objetivos e Estrutura

Esta dissertação tem como principal objetivo desenvolver uma análise exploratória baseada nos SIG e em IA para recolha e tratamento de informação histórica e analise dos padrões agroecológicos e especificidades agrícolas regionais em Portugal durante a década de 1950. Para atingir este objetivo esta dissertação propõe uma abordagem interdisciplinar que combina métodos tradicionais de análise de dados com técnicas avançadas de inteligência artificial incluindo a implementação de *Text mining* e *machine learning*.

De seguida apresentam-se os sub-objetivos:

1- Desenvolver uma abordagem exploratória com recurso à IA e aos SIG para recolha e tratamento da informação histórica textual e cartográfica para um formato estruturado adequado para análise espacial;
2- Explorar algoritmos de *Natural Language Processing* e *Machine Learning* para averiguar as especificidades agrícolas regionais em Portugal e aferir de que forma as mesmas se agrupam entre diferentes regiões;
3- Criar e disponibilizar um mapa agrícola da década de 1950 com base na informação histórica cartográfica.

Deste modo, elaboraram-se as seguintes hipóteses de investigação:

Hipótese 1 – A IA melhora o processo de recolha e tratamento de informação histórica em formatos textual e cartográfico, servindo como um reforço importante para as ferramentas já existentes;

Hipótese 2 – Os algoritmos de *Natural Language Processing* e *Machine Learning* permitem clusterizar especificidades agrícolas regionais;

Hipótese 3- A importância das culturas agrícolas varia entre as diferentes regiões da área de estudo.

Para atender aos objetivos estabelecidos e verificar as hipóteses mencionadas anteriormente, a dissertação foi estruturada em cinco capítulos. No Capítulo I, é introduzida a temática, justificando a sua importância e relevância. Serão igualmente apresentados os objetivos e as hipóteses de investigação. O Capítulo II inicia com a exposição do tema das humanidades digitais e a apresentação dos SIG como uma ferramenta de apoio a este campo de estudo. Em seguida, faz-se o enquadramento conceptual de outras ferramentas amplamente utilizadas nas humanidades digitais que complementam e auxiliam a abordagem metodológica proposta para responder às hipóteses de investigação. Este capítulo também aborda o tema da IA, incluindo conceitos inerentes como Big Data, Data Mining e Machine Learning. O Capítulo III apresenta a metodologia desenvolvida para o tratamento dos dados, incluindo uma descrição da principal fonte de informação utilizada para o estudo, seguida dos métodos e técnicas selecionados para o tratamento da informação e análise espacial. O Capítulo IV apresenta os resultados e a discussão. O capítulo V finaliza com as principais conclusões e sugestões de estudos futuros incluindo propostas de melhoria da metodologia.

Capítulo 2 – Inovação, Inteligência Artificial e os Sistemas de Informação Geográfica

2.1. O alcance da Inteligência Artificial

Vivemos atualmente na era da inovação tecnológica, frequentemente referida como a "4ª Revolução Industrial" (Lu, 2017). Esta fase caracteriza-se por um avanço exponencial no desenvolvimento industrial, particularmente no setor tecnológico, reconfigurando, assim, os padrões sociais (Maynard, 2015). Esta revolução, marcada por uma mudança paradigmática na forma como interagimos e percebemos o mundo, tem sido impulsionada pela crescente interconexão entre indivíduos e dispositivos, tornando os processos quotidianos progressivamente mais automatizados e a vida, por conseguinte, aparentemente mais regida pela automação. Este fenómeno tem impactado substancialmente diversas indústrias ao redor do globo.

Conforme Sarker (2022) salienta, as três componentes fulcrais desta nova fase tecnológica emergente englobam: (i) *Automation*, que visa reduzir progressivamente o envolvimento humano nas operações; (ii) *Intelligent*, ressaltando a capacidade de extrair informações cruciais e aplicar esse conhecimento de forma eficaz; (iii) o *smart computing*, onde análises e relatórios têm desempenhado um papel crucial no desenvolvimento de novas tecnologias. Neste contexto, a IA destaca-se como uma tecnologia de vanguarda, sendo um vetor primordial de investigação em campos tão diversos como engenharia, ciência, educação, medicina, comércio e finanças (Halal, 2003).

A IA, apesar de não possuir uma definição universalmente consensual, é geralmente associada à capacidade de um sistema aprender a partir da experiência e do ajuste de *inputs* (Duan, 2019). Atualmente, admite-se a possibilidade de desenvolver sistemas capazes de superar a eficiência humana em diversas tarefas (Duan & Edwards, 2019). Predominantemente, a IA é percebida como uma ferramenta decisiva, assumindo um papel crítico na história da tecnologia. A sua aplicação na tomada de decisões pode ser categorizada em dois aspectos: o apoio e assistência ao ser humano e, alternativamente, um papel mais autónomo, onde a máquina toma decisões de forma independente, suplantando a intervenção humana (Edwards *et al.*, 2000).

Alan Turing é considerado o pai da Inteligência artificial tendo contribuído com dois estudos importantes, e.g., a máquina de Turing em 1936 que é considerada como a fase

inicial dos computadores modernos. Já os neuro-fisiologistas W. McCulloch e W. Pitts estabelecem em 1943 o primeiro modelo matemático de redes neuronais fundamentado na biologia neuronal e sendo amplamente reconhecido como o precursor das redes neuronais artificiais. Em 1949, surgiu a teoria do "*Hebbian learning*", formulada por Hebb, sendo um mecanismo de aprendizagem baseado na neuropsicologia que operava através da análise estatística de dados e da exploração das relações entre amostras (Kuriscak, 2019). Esta teoria, considerada a precursora da *Machine Learning*, baseia-se em princípios análogos aos processos cognitivos humanos (Jiang, 2022). Em 1950, Turing, apresentou o "Teste de Turing" como um método para avaliar a inteligência de uma máquina. De acordo com este teste, se uma máquina conseguir imitar um ser humano numa conversa escrita, pode ser considerada inteligente. Este teste continua a ser uma referência nas discussões sobre inteligência artificial (Turing, 1950; Copeland, 2000). Em 1956, durante o *Dortmouth Summer Research Project on Artificial Intelligence*, John McCarthy define o termo "Inteligência Artificial" referindo-se a qualquer aspecto de aprendizagem ou função inteligente pode ser tão precisamente descrito que uma máquina seja capaz de simular tal comportamento (McCarthy *et al.*, 1956). Atualmente, a IA demonstra uma capacidade única de analisar e adaptar-se a contextos variados, tornando-se uma ferramenta indispensável em diversos ramos da ciência. Contudo, face às dinâmicas imprevisíveis do mundo real, continua a ser desafiador criar modelos eficazes que incorporem todos os dados necessários para refletir a realidade o mais fielmente possível.

Apesar deste reconhecimento, a IA enfrentou um período de desafios significativos ao longo do seu desenvolvimento, marcados por obstáculos fundamentais como a falta de ferramentas algorítmicas avançadas e a limitada capacidade eletrónica e computacional. Estes fatores necessitavam de investimentos de grande escala, resultando num progresso esporádico e lento (Jiang *et al.*, 2022). Após este período de estagnação, que ocorreram entre as décadas de 1970 e 1980, a IA ressurgiu como uma ferramenta de relevância primordial, impulsionada em grande parte pelos êxitos alcançados com algoritmos de *machine learning*. No final dos anos 1980, emergiram diversos modelos inovadores que marcaram a área, destacando-se a árvore de decisão proposta por J.R. Quinlan em 1986, a máquina de vetores de suporte de Vapnik e Cortes em 1995, e a *Random Forest* de Breiman em 2001 (Jiang *et al.*, 2022).

Atualmente, a IA pode ser categorizada em cinco vertentes principais: Analítica, Funcional, Interativa, Textual e Visual (Figura 1) (Sarker, 2022). A IA analítica foca-se na identificação, interpretação e comunicação de padrões existentes nos dados de entrada. O objetivo principal é a descoberta de novos conhecimentos, baseando-se nos padrões e relações identificadas nos dados, a fim de facilitar uma tomada de decisão eficaz e informada. Neste âmbito, têm sido desenvolvidas várias técnicas de *machine learning* e *deep learning* para abordar problemas do mundo real e sugerir as melhores alternativas (Ertel, 2018). A IA funcional, embora similar à analítica, procura por padrões e relações nos dados, e distingue-se pela sua capacidade de tomar decisões e executar ações, ao invés de apenas recomendar soluções. Está frequentemente associada a aplicações no campo da robótica. Por sua vez, a IA interativa relaciona-se com a comunicação interativa e automática, estando já integrada em muitos dos dispositivos do nosso quotidiano como smartphones, tablets, etc. Adicionalmente, a IA textual, com a habilidade de reconhecer e processar texto, é capaz de gerar respostas e conteúdo relevante. Por fim, a IA visual especializa-se na classificação de objetos e na conversão de imagens ou vídeos em dados analisáveis (Sarker, 2021).

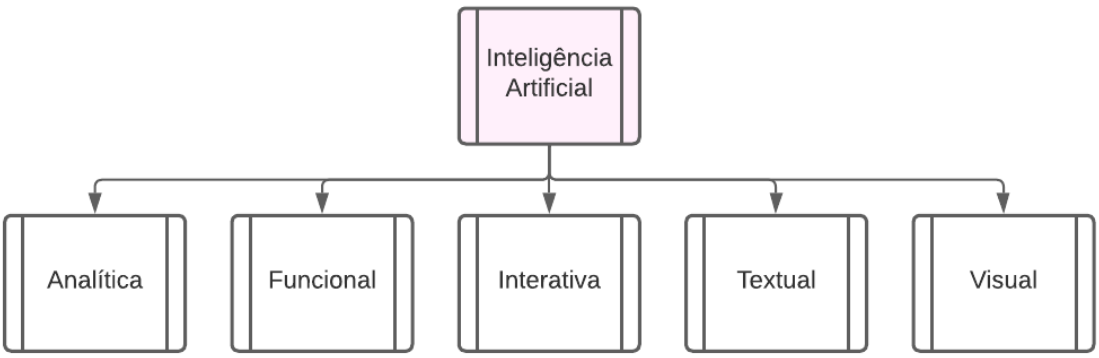


Figura 1 - Divisões da Inteligência Artificial. (Fonte: elaboração própria)

2.2. A Era do *Big Data* e a Revolução dos Dados

Atualmente, com a emergência da Internet das Coisas (IoT), observa-se um incremento exponencial no volume de dados gerados. Este fenómeno originou uma transformação paradigmática, inserindo-nos num contexto onde a geração e o acesso a dados são facilitados. Estes dados têm sido extensivamente explorados em diversas áreas, destacando-se a segurança alimentar e a gestão agrícola, com o intuito de otimizar

recursos e promover a sua sustentabilidade (Misra *et al.*, 2022). O desenvolvimento das IoT e da economia digital situou a sociedade contemporânea na era do *big data*, particularmente devido ao aumento exponencial de dados gerados e partilhados por empresas, indivíduos e entidades públicas. Segundo Dobre e Xhafa (2014), estima-se que diariamente sejam produzidos cerca de 2.5 quintilhões de bytes de dados. Com o crescente volume de dados, associado à partilha e aplicação conjunta de técnicas de inteligência artificial para analisar vastas quantidades de dados, muitos dos quais são não estruturados, torna-se possível obter novo conhecimento (Chang *et al.*, 2021).

O conceito de *big data* foi introduzido em 1997, num artigo publicado por investigadores da NASA, os quais o associaram a grandes conjuntos de dados que desafiam a capacidade de memória. Num relatório publicado em 2011 pela empresa McKinsey, o *big data* foi reconhecido como a fronteira para a inovação, competitividade e produtividade (Côrte-Real, 2022; Brown *et al.*, 2011). O avanço tecnológico permitiu que o *big data* ganhasse uma importância cada vez maior tanto não só no meio académico, mas também empresarial (Fichman *et al.*, 2014; Chen *et al.*, 2015). De acordo com Côrte-Real (2022, pp. 19), "Atualmente, as empresas que investem em ferramentas de *Big data* [...] tendem a ter ganhos de produtividade de 5% a 6% de lucro acima dos concorrentes".

O *big data* é caracterizado pela combinação dos 4 "v's": velocidade, volume, variedade e veracidade. A velocidade refere-se à rapidez com que a informação é produzida, acredita-se que desde 2018, 90% da informação mundial foi criada (Côrte-Real, 2022, pp. 23). Outra característica fundamental é o volume (Najafabadi *et al*, 2015). O vasto volume de dados adquiridos representa um desafio para a computação convencional, uma vez que requer estratégias eficazes de armazenamento e de consulta para realizar as análises pretendidas. O *big data* também é variedade, dado que incorpora dados estruturados e semi-estruturados, sendo que aproximadamente 20% destes dados são estruturados e 80% são não estruturados (como registos de web, mensagens e imagens) (Côrte-Real, 2022). O último "v" diz respeito à veracidade, relacionando-se com a qualidade e confiança nos dados. Os dados não estruturados, por norma, possuem bastante ruído, sendo, portanto, importante estabelecer uma estratégia para o controlo de qualidade desses dados, através de regras de exclusão de determinados dados (Côrte-Real, 2022).

O impacto que o *big data* tem tido em praticamente todo o domínio comercial e científico proporcionou à comunidade de geocientistas uma evolução no seu campo de estudo, transacionando da fraca qualidade de dados para uma rica e volumosa quantidade de dados (Karpatne *et al.*, 2018). Este avanço apenas foi possível com a evolução dos sistemas de observação terrestre (deteção remota), a melhoria das capacidades de computação que atualmente são capazes de realizar simulações em grande escala do sistema terrestre e, sobretudo, na democratização dos dados de Internet que podem ser adquiridos, guardados e processados em ambientes de *cloud* (Karpatne *et al.*, 2018). Por exemplo, o programa de satélite *Landsat*, que teve início na década de 70, adquiria dados em níveis que excediam as capacidades analíticas computacionais da época. As melhorias subsequentes no equipamento, com sensores de maior resolução, significam que mesmo passados 50 anos, ainda é um desafio lidar com todo este volume de dados, mesmo para os sistemas mais poderosos. Contudo, o volume não é a única característica que distingue os dados atuais; atualmente, esta informação é recolhida através de diversas fontes, como redes sociais, crowdsourcing, câmaras, entre outras. Com o *big data*, são feitas previsões com base em data mining para encontrar padrões e correlações entre toda a informação adquirida. Como mencionado anteriormente, o *big data* não se trata apenas do volume, mas também da variedade das diferentes fontes de dados, bem como da velocidade com que toda a informação pode ser analisada (Harvey e Goodchild, 2015).

2.3. Os SIG e as Humanidades Digitais

Segundo Harvey e Goodchild (2015), a maior transformação da ciência baseada em dados na geografia foi a capacidade de explorar dados e gerar hipóteses. Os dados espaciais contribuem para classificar o contexto ambiental e geográfico, facilitando etapas importantes do processo científico, como a seleção de locais de estudo, o reconhecimento, a etnografia, o design experimental e a logística. Na última década, com o avanço das tecnologias GPS e a partilha de informação com componente espacial, tornou-se mais fácil criar dados de alta qualidade com atributos geográficos. O progresso tecnológico e da informação tem sido um motor para a geração de dados espaciais, permitindo, por exemplo, localizar pessoas através dos seus dispositivos móveis e captar numerosas informações sobre os indivíduos, incluindo suas relações e locais frequentados. Estes dados têm aplicação em diversas áreas, como análises culturais e sociais, e até para fins comerciais.

Atualmente, tanto a indústria privada quanto o setor público mostram interesse em criar e obter dados com componente geográfica (Mennis & Guo, 2009). O crescimento destes dados torna os SIG um sistema essencial para representar tais interações (Goyal et al., 2017), pois permitem a importação, armazenamento, gestão e exportação de dados espaciais. No entanto, os SIG vão além do mapeamento, devendo ser vistos como tecnologias de bases de dados, onde cada componente de dados—seja uma linha estatística, um campo de texto ou uma imagem—está associada a uma componente espacial por meio da representação gráfica das coordenadas (Gregory & Healey, 2007). Um exemplo disso foi a revolução provocada pelos SIG nas técnicas utilizadas por alguns campos de investigação (Gregory, 2001). Por exemplo, na disciplina da História, os historiadores rapidamente reconheceram o vasto potencial e o valor acrescentado dos SIG nos seus trabalhos, o que levou à consolidação do termo "Historical GIS" (HGIS) (Gregory & Ell, 2007).

Knowles (2005) define HGIS como a aplicação dos SIG, juntamente com outras tecnologias geoespaciais, no estudo da História. O primeiro estudo significativo nesta área foi "*The Great Britain Historical GIS Project: From Maps to Changing Human Geography*" (Gregory *et al.*, 2002), que visava criar uma base de dados cartográfica das alterações dos limites administrativos do Reino Unido. Este estudo foi pioneiro no desenvolvimento dos métodos de HGIS, introduzindo abordagens inovadoras que combinavam mapas e informações textuais para verificar, com precisão, as mudanças nos limites administrativos. Além disso, promoveu o desenvolvimento de metodologias para integrar as informações recolhidas em ambiente SIG. Gregory (2001) destaca três vantagens do uso dos SIG em pesquisas históricas: i) a capacidade de localizar espacialmente informações das fontes históricas para construir bases de dados estruturadas; ii) a possibilidade de visualizar informações por meio de mapas ou tecnologias de animação de paisagens virtuais; e iii) a habilidade de realizar análises espaciais onde a localização das variáveis é um elemento-chave. Relativamente às bases de dados, estas são compostas por uma variedade de fontes históricas integradas, com o objetivo de superar as limitações de cada fonte individual (Gregory & Healey, 2007). Contudo, a construção de uma base de dados pode tornar-se uma das etapas mais desafiantes num projeto de SIG (Knowles, 1999). Outras disciplinas, como a Arqueologia e a Literatura, também têm utilizado os SIG para incorporar a dimensão espacial dos acontecimentos, permitindo a integração de informações específicas destas disciplinas em

dados quantitativos (Connolly & Lake, 2006). Por exemplo, a apresentação de dados em formato visual, como mapas, permitiu adquirir novos conhecimentos e formular ou confirmar novas hipóteses em diversas áreas de investigação (Keim *et al.*, 2004).

A procura de respostas e desenvolvimento de novas narrativas com foco exclusivo na tecnologia e nos dados contribui para o surgimento do campo interdisciplinar das "Humanidades Digitais" que engloba sub-campos como o HGIS, as *Spatial Humanities*, etc. A sua natureza interdisciplinar, integrando métodos e ferramentas de diversas áreas, como história, literatura, linguística, geografia, ciências da computação e ciências da informação permite explorar novas perspetivas e abordar questões complexas de forma inovadora (Bodenhamer *et al.* 2010; Gregory, 2014). Este novo campo de estudo envolve várias tarefas e são diversas as ferramentas que podem ser utilizadas, sejam elas computacionais, estatísticas ou visuais, muitas vezes numa combinação de todas elas (e.g., Linguística de Corpus; Processamento de Linguagem Natural (PLN); Padrões Textuais; Representações Espaciais Qualitativas, etc.) (Mennis & Guo, 2009). A Figura 2 apresenta as diversas áreas de investigação nas Humanidades Digitais, delineando os seus inter-relacionamento.

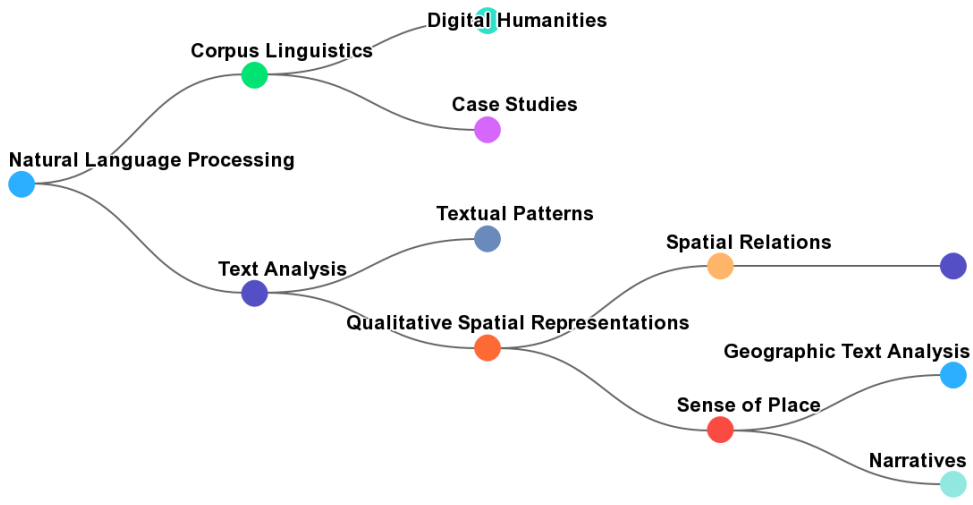


Figura 2 – Principais áreas de investigação nas Humanidades Digitais. Fonte: Scopus

2.4. Métodos, Tecnologias e ferramentas das Humanidades Digitais

2.4.1. Reconhecimento Ótico de Caracteres (OCR)

O Reconhecimento Ótico de Caracteres (OCR) tem-se revelado como uma das tecnologias mais promissoras nas áreas das ciências sociais e das humanidades, conforme indicado por Hegghammer (2022). Esta tecnologia, que mimetiza a capacidade humana de ler e processar informação, tem evoluído significativamente desde a sua conceção (Mithe *et al.*, 2013). O aparecimento do OCR remonta à invenção do scanner de retina, o qual utilizava um sistema de transmissão de imagens baseado num mosaico de fotocélulas (Mantas, 1986). Inicialmente concebido como uma ferramenta auxiliar para pessoas com deficiência visual, o OCR rapidamente transformou-se num campo de pesquisa e desenvolvimento (Berchmans & Kumar, 2014).

Berchmans e Kumar (2014) identificam quatro gerações distintas de produtos OCR. A primeira, surgida na década de 1960, limitava-se ao reconhecimento de um conjunto restrito de fontes e formas de caracteres (Mori *et al.*, 1992). A segunda geração, estendendo-se de 1965 até meados dos anos 1970, notabilizou-se pela capacidade de reconhecer tanto caracteres impressos quanto manuscritos, embora focada principalmente em dígitos numéricos. A terceira geração, entre 1975 e 1985, ganhou popularidade com o avanço da capacidade computacional, possibilitando o processamento de dados em maior escala (Nagy, 1992). A quarta e mais recente geração representa um salto qualitativo, com softwares capazes de reconhecer uma ampla gama de caracteres, incluindo símbolos matemáticos e alfabetos diversos como o Japonês, Chinês, Árabe e Romano, mesmo em documentos com ruído visual (Berchmans & Kumar, 2014).

Os softwares de OCR contemporâneos procuram simular as funções óticas e cognitivas humanas, inserindo-se no âmbito da inteligência artificial. Apesar de sua eficácia no reconhecimento de caracteres, estas ferramentas ainda enfrentam desafios, sobretudo quando comparadas à capacidade humana de interpretar textos sob diversas condições (Holley, 2009). No campo da História, o OCR é frequentemente utilizado no tratamento de fontes históricas, embora estas apresentem desafios específicos devido à sua condição física, como páginas com ruído visual, rasgos, manchas de humidade, tinta borrada e baixo contraste (Koistinen *et al.*, 2017; Thompson *et al.*, 2015).

Para mitigar esses desafios, Holley (2009) e Volk, Furrer e Sennrich (2011) sugerem uma série de práticas a serem adotadas no tratamento inicial de fontes históricas

de baixa qualidade visual. Entre as recomendações estão a utilização de uma resolução mínima de 300dpi na digitalização da fonte original e a melhoria do contraste da imagem, frequentemente alcançada através de um processo de binarização que converte as cores em preto e branco, funcionalidade já incorporada em muitos softwares OCR atuais. Estas práticas visam elevar a precisão dos resultados obtidos através do OCR, especialmente em documentos históricos com desafios específicos de preservação e legibilidade.

2.4.2. Spatial Data mining

A capacidade de adquirir, processar, partilhar e analisar grandes volumes de informação assinala uma nova fase na sociedade contemporânea. Atualmente, entramos na quarta fase de desenvolvimento científico, que além dos métodos tradicionais de observação individual, desenvolvimento de teorias ou simulações computacionais, também se fundamenta numa nova abundância de dados digitais (Hey *et al.*, 2009). No campo da Ciência da Informação Geográfica (CIG), a comunidade científica considera a *Data Mining* um caminho promissor para melhorar a qualidade da análise espacial. No entanto, nas últimas décadas, a Geografia já utilizava métodos subjacentes à *Data Mining* (Bação & Painho, 2003).

A *Data mining* consiste em melhorar a ligação entre uma base de dados e os humanos, de modo que toda a informação seja explorada e modelada, algo que a Geografia tem realizado ao longo dos anos. Um exemplo disto é o estudo realizado pelo Dr. John Snow durante a epidemia de cólera em Londres, em 1854, que fez uma descoberta de grande importância para a época com recurso à análise espacial, muito antes da utilização generalizada dos Sistemas de Informação Geográfica (Bação & Painho, 2003). Nas últimas décadas, foram desenvolvidos diversos trabalhos com análise espacial e padrões geográficos utilizando técnicas de data mining como o clustering, criação de modelos preditivos para aplicações agrícolas e ambientais treinados com grandes volumes de dados ou integração de informação geográfica com data mining de forma a mapear a distribuição espacial de doenças e fatores de risco (Shekhar e Huang, 2001; Chlingaryan *et al*., 2018, Nuckols et al., 2004).

2.4.3. Natural Language Processing

A *Natural Language Processing* (NLP) emprega técnicas computacionais que permite às máquinas analisar, compreender e gerar linguagem humana (Côrte-Real, 2022). De acordo com a definição proposta por Liddy (2001) a *Natural language*

processing (NLP) é caracterizada pelo conjunto de técnicas computacionais para a análise e representação de semânticas textuais a um ou mais níveis linguísticos. Tem o propósito de chegar a uma abordagem humana relativamente ao processamento da linguagem para a realização e aplicação de tarefas. A NLP é um campo científico onde as ciências computacionais, a linguagem e a matemática interagem em conjunto com o objetivo de transformar a linguagem humana em comandos que possam ser executados por um computador.

O desenvolvimento da NLP é dividido em 4 diferentes períodos segundo (Kang *et al.*, 2020): período de "germinação" antes do ano de 1956; o período de desenvolvimento rápido entre 1957 e 1970; o período de desenvolvimento lento entre 1971 e 1993; o período de recuperação desde 1994 até ao momento atual. Especificamente, decorria o ano de 1948 quando Shannon aplicou o modelo probabilístico de Markov de forma a automatizar a linguagem descritiva, ao aplicar um algoritmo probabilístico do processamento de linguagem, conseguiu medir a quantidade de informação que a linguagem humana continha (Kang et al, 2020). Em 1954, Georgetown-IBM apresentou a primeira máquina de tradução que foi capaz de traduzir cerca de 60 frases da linguagem russa para a inglesa, este passo foi marcante na era de desenvolvimento da NLP (Krutilla *et al.*, 2022). Após este marco, a NLP evolui exponencialmente, uma vez que passa a integrar o campo de estudo da IA. Nesta década Chomsky, começou por estudar teorias da linguagem formal e a utilizar métodos probabilísticos que evoluíam cada vez mais à medida que o progresso das duas áreas era feito.

Atualmente, existem duas subdivisões principais da NLP, a *Natural Language Understanding* (NLU) e a *Natural Language Generation* (NLG) (Kang et al, 2020). Bates (1995) entende que a *Natural Language Understanding* é um ponto fulcral para que os computadores lidem com uma das características mais complexas do ser humano, a linguagem. A linguagem é um aspeto fundamental na espécie humana, o uso fluente da mesma é muitas vezes visto como uma característica de inteligência, deste modo verifica-se a razão para a importância de conseguir que os computadores a processem. Assim, o principal objetivo da NLU passa por compreender a linguagem humana, por exemplo, retirando informação de documentos. Por outro lado, a *Natural Language Generation*, pode ser definida como a capacidade de gerar textos através de dados computacionais. A capacidade de criar textos pode realizar relatórios, até criar um discurso. A utilização da NLG implica a passagem por três diferentes fases que são divididas da seguinte forma:

(a) Organização de Texto: Implica a definição do conteúdo básico do texto, onde são utilizadas teorias de organização de texto para a ordenação da informação de forma coerente; (b) Organização de Frases: Nesta fase a informação é estruturada em frases e parágrafos, sendo que é aqui que se utiliza adequadamente as conjunções e preposições; (c) Realização: No terceiro e último processo são realizadas a correções gramaticais. (Cawsey *et al.*,1997). Em suma, a NLG implica a produção de texto numa linguagem natural, programada por seres-humanos baseada em processos de estruturação de dados, textos, gráficos (McDonald, 2010).

O funcionamento da NLP implica a utilização de cinco fases de compreensão, à semelhança do processo de compreensão que o ser-humano utiliza, estas fases englobam 1) fonologia; 2) morfologia; 3) léxico; 4) Semântica e 5) Sintaxe (Liddy, 2001):

1. Fonologia: É o estudo do som, a sua estrutura e funcionamento, procura perceber como as frases e palavras estão organizadas (Bisol, 2005);

2. Morfologia: Lida com a estrutura da palavra, onde se pode ter em conta o prefixo, e o sufixo. Neste sentido a NLP tem a capacidade de compreender o sentido da palavra com base no prefixo ou sufixo e a capacidade de formar novas palavras (Liddy, 2001; Eliseu, 2014);

3. Léxico: A compreensão do Léxico na componente da NLP desempenha o papel da interpretação do significado das palavras (Biderman, 1996);

4. Semântica: Este nível de estudo permite determinar o significado da frase com base nas palavras que compõe a mesma (Liddy, 2001);

5. Sintaxe: É o conjunto de regras que permite associar o som ao significado (Eliseu, 2014).

O progresso da IA veio acelerar o desenvolvimento da NLP e atualmente esta ferramenta vai além da capacidade de uma máquina em traduzir textos, tornando-se possível reconhecer padrões textuais, realizar análises de sentimento e reconhecer um discurso automaticamente. Assim, a NLP é um campo multidisciplinar que congrega áreas das ciências da computação, linguística e inteligência artificial. A NLP pode ter uma variedade de aplicações, nas quais são empregues diferentes técnicas conforme o objetivo específico. Algumas das aplicações da NLP incluem:

1) Pesquisa Semântica: Destina-se a compreender a intenção, contexto e relação entre palavras. Está relacionada com a contextualização e a pesquisa de frases em bases de dados, visando extrair significados semânticos relevantes;

2) Tradução Automática: Composta por duas fases distintas, a análise da linguagem de origem e a geração da linguagem de destino. O computador analisa a estrutura do texto original, segmentando-o em unidades linguísticas que podem ser traduzidas individualmente. Esta aplicação é capaz de traduzir grandes volumes de texto em um curto período;

3) Sumarização de Texto: Amplamente utilizada em contextos académicos, envolve técnicas de extração e abstração de informações relevantes de um texto. Dada a vasta quantidade de informações disponíveis na Internet, este método tem sido aprimorado continuamente, visando criar uma síntese comunicativa do conteúdo original. Motores de busca, por exemplo, podem fornecer uma visão sumária rápida de documentos extensos (Alomari *et al.* 2021);

4) Análise de Sentimento: Considerada a abordagem mais reconhecida e aplicada da NLP, visa compreender os sentimentos expressos por indivíduos com base em opiniões recolhidas de questionários, artigos de opinião ou conteúdos de redes sociais. O resultado básico dessa análise é a classificação dos sentimentos em três escalas distintas: positiva, negativa e neutra (ABRO *et al.*, 2023).

Estas funcionalidades permitem automatizar rotinas de consulta de sites ou documentos produzidos (Côrte-Real, 2022).

2.4.4. Text mining

O avanço tecnológico tem propiciado a criação de volumes crescentes de dados, particularmente em formato textual, manifestando-se em plataformas como fóruns de discussão, mensagens de texto, redes sociais e publicações científicas (Ferreira-Mello *et al.*, 2019). Neste cenário, o texto tornou-se o principal meio de comunicação e troca de informações, evidenciando a necessidade de desenvolver tecnologias que permitam extrair informações relevantes destes vastos conjuntos de dados para análises diversas (Witten, 2004). Estamos, portanto, na era do *big data*, onde cerca de 80% dos dados não estão estruturados numa forma de linguagem natural como livros, notícias, artigos

científicos, redes sociais. *Text Mining* é um processo conjunto entre a *natural language processing*, *data mining*, *machine learning* e estatística, que permite formular novo conhecimento deste texto não estruturado (Zong *et al.*,2021).

Inicialmente os académicos utilizavam uma abordagem qualitativa para analisar o texto, no entanto esta tipo de análise, para além de consumir tempo, era um trabalho intensivo que lidava com os desafios relacionados com a interpretação (Kobayasgi *et al.*,2018; Jamiy *et al.*, 2015). Desde a última década que se tem desenvolvido funcionalidades de automatização destas análises de texto (Wiedeman, 2013). *Text Mining* e a NLP tornam-se essenciais quando o corpo dos textos é extenso. A *Text mining* faz parte de um conjunto de técnica de IA que converte dados não estruturados para estruturados, onde se utiliza a NLP para melhorar o algoritmo de *machine learning* para as análises.

Hotho, Nürnberger e Paaß (2005) apontam que o conceito de *Text mining* foi inicialmente proposto no artigo "*Knowledge Discovery in Textual Databases*" de Feldman e Dagan (1995). Este trabalho procurou estabelecer métodos alternativos aos existentes, que se mostravam insuficientes perante a crescente produção de informação. A abordagem sugerida consistia na combinação da descoberta de conhecimento em bases de dados – focada na exploração e processamento de grandes volumes de dados à procura de padrões relevantes – com a análise de informação textual. Este processo, contudo, exigia uma estruturação prévia dos dados. Atualmente, a *Text mining* é reconhecida como uma das ferramentas mais promissoras, particularmente com os avanços recentes da IA. Ansari *et al.* (2021) destacam o imenso potencial de utilizar dados não estruturados, transformando-os para análises significativas. Gandomi e Haider (2015) ilustram o potencial desta ferramenta onde a *Text mining* se enquadra na esfera das análises de *Big data*. Esta área engloba a aplicação de análise estatística, linguagem computacional e *Machine learning* para a interpretação e extração de informações valiosas a partir de grandes volumes de texto. Nos últimos anos, tem-se desenvolvido diversas abordagens para explorar informação espacial, temporal e encontrar padrões na informação presente nos textos (Almatar *et al.*, 2020). Uma vez que a *Text mining* utilizada sobretudo técnicas de predição, clusterização e análises de tendência torna-se possível utilizar esta ferramenta em conjunto com dados com atributos geográficos.

Durante o processo de *Text mining*, um dos aspetos fundamentais é o pré-processamento do texto. O pré-processamento pode tornar o modelo inconclusivo, pois pode levar a resultados influenciados pela incapacidade dos dados de atenderem ao propósito do modelo. Esta fase permite transformar frases e palavras em *tokens*, utilizando diversas técnicas para uniformizar o texto, como a conversão de todas as palavras para minúsculas, a remoção de palavras irrelevantes para o problema, a correção de erros ortográficos, a stemização e a lematização, entre outras (Hickman *et al.*, 2022).

A tokenização é o processo elementar do pré-processamento de texto, que envolve a segmentação do texto, ou seja, a separação de frases e palavras em unidades significativas para a análise. A divisão do conteúdo original do documento em elementos singulares permite representar os dados na forma de vetores no espaço, possibilitando a obtenção do peso de cada elemento com base na frequência em que aparece no texto. Assim, é possível obter uma identificação única das palavras na sua forma básica, além de criar grupos isolados de *tokens* que, em conjunto, podem agregar mais valor à análise. A primeira parte da tokenização é a conversão do documento em contagem de palavras; a segunda operação envolve a remoção de espaços; e, por último, o conteúdo é transformado em tokens (Uysal *et al.*, 2014; Hickman *et al.*, 2022). A remoção de *stopwords* é, geralmente, o segundo tratamento aplicado ao texto. As *stopwords* são divisões da linguagem natural que devem ser removidas para eliminar ruído no texto. Essas *stopwords* são, na sua maioria, determinantes, preposições e pronomes (Vijayarani *et al.*, 2015). Também é possível criar uma lista de *stopwords* verificando os termos mais frequentes no documento e avaliando se fazem sentido para a análise (Uysal *et al.*, 2014). O processo de stemização procura construir o formato básico das palavras retirando afixos, plurais e terminações de verbos (Hotho *et al.*, 2005). Por sua vez, o processo de lematização envolve transformar os verbos para o infinitivo e os nomes para a sua forma singular. No entanto, para o fazer, é necessário conhecer a forma da palavra através das partes do discurso de cada palavra no documento, que devem corresponder.

Depois de preparar todos os dados, é necessário converter os dados textuais em vetores numéricos, uma vez que a maioria dos algoritmos de aprendizagem automática trabalham com inputs numéricos (vetorização) (Abubakar *et al.*, 2022). A máquina processa um arranjo numérico numa matriz de duas dimensões em que o objeto definido pela classe é representado numa linha e a característica numa coluna (Rani *et al.*, 2022). A vetorização é especialmente importante para o pré-processamento de texto, pois mapeia

diretamente as palavras para uma representação numérica, facilitando a análise (Danyal *et al.*, 2024). Algumas das técnicas mais comumente utilizadas são: bag of words, word2vec, doc2vec e Term Frequency-Inverse Document Frequency (TF-IDF) (Abubakar *et al.*, 2022), as quais são descritas a seguir.

Bag of Words: Modelo utilizado principalmente para categorização de texto. Este modelo aprende com o vocabulário fornecido e conta o número de vezes que cada palavra surge num documento. É relativamente simples; no entanto, é aplicável apenas quando a independência de cada palavra é garantida. A ordem das palavras não é considerada, apenas a frequência com que aparecem, o que torna este método bastante prático (Rani *et al*., 2022; Bhansali *et al*., 2022).

Word2Vec: Trata-se de um modelo não supervisionado baseado em redes neuronais. O treino deste modelo permite associar palavras com o mesmo significado semântico; palavras como "feliz" e "contente" passam a ser representadas com vetores relativamente próximos no espaço. No entanto, trata-se de um modelo "blackbox", uma vez que todo o treino é feito de forma interna e as relações semânticas encontradas baseiam-se nos inputs fornecidos. Com uma dimensão elevada, a representação de vetores baseados na proximidade semântica torna-se mais enriquecedora (Rani et al., 2022).

Doc2Vec: Foi criado por Le e Mikolov (2014) como uma extensão do word2vec. Este modelo utiliza redes neuronais para treinar frases, enquanto gera um vetor correspondente a cada palavra. O modelo assume que a ocorrência de cada palavra está relacionada apenas com um número específico de palavras que a antecedem (Weiwei et al., 2024).

Term Frequency-Inverse Document Frequency (TF-IDF): Método que determina a importância relativa das palavras num documento através da proporção inversa das palavras sobre todo o corpo do documento. É composto por dois elementos fundamentais: TF – frequência do termo e IDF – inverso da frequência do termo no documento (Trstenjak et al., 2014). Dada uma coleção de documentos (D), uma palavra (w) e um documento individual (d), o cálculo efetua-se da seguinte forma:

$$wd = fw, d * \log(|D|/fw, D) \quad (1)$$

Onde $f_{w,d}$ são o número de vezes que o w aparece no d, e |D| é o tamanho do corpo do documento e f_{wD} é igual ao número de documentos onde o w aparece no (D) (Salton e Buckley,1988; Berger *et al.*, 2000).

2.4.5. Machine Learning

Machine Learning faz parte do ramo da IA para o desenvolvimento de algoritmos de aprendizagem com base em dados e que procura encontrar e aprender padrões e relações com base nos dados. Na última década, estes algoritmos têm vindo a ganhar cada vez mais importância em diferentes domínios sociais e do quotidiano revolucionando muitos aspetos da sociedade atual (e.g., comércio, transportes, entretenimento, etc. (Karpatne *et al.*, 2018). Por exemplo, a simples correção de uma palavra numa mensagem enviada através de um smartphone ou a sugestão de produtos com base nas nossas pesquisas, são aspetos que fazem parte deste campo.

Os algoritmos de machine learning são aproximadores universais que funcionam como um método para encontrar padrões e relações presentes num sistema através da utilização de um conjunto de dados de treino (Lary,2010). O desempenho de um modelo de machine learning é medido por uma métrica de performance que evolui com a experiência ao longo de diversas corridas. Para calcular a performance de um modelo ou algoritmo, são utilizados diversos modelos estatísticos e matemáticos. No fim do processo de aprendizagem o modelo treinado pode ser utilizado para classificar, prever ou agrupar (clustering) novos conjuntos de dados utilizando esse mesmo modelo. Usualmente classifica-se a machine learning em diferentes tipo de categoria baseada no tipo de aprendizagem (supervisionado ou não-supervisionada) e modelos de aprendizagem (classificação, regressão, clustering e redução da dimensionalidade) (Liakos *et al.*, 2018). Entende-se por métodos supervisionados aqueles que como o nome indica têm um supervisor, isto é, algo que instrui o sistema com exemplos de treino. Estes modelos utilizam dados treinados para criar o algoritmo e são posteriormente aplicados a outro conjunto de dados desconhecidos (Cunningham *et al.*, 2008). Em contrapartida os métodos não-supervisionados podem ser descritos como métodos que procuram identificar padrões e estruturas em conjuntos de dados sem necessitar de dados previamente treinados. Por norma são utilizados para agrupar dados semelhantes e reduzir a sua dimensionalidade (Bishop & Nasrabadi, 2006). Alguns autores consideram a Machine learning o futuro da computação moderna pois muitas das suas técnicas já se estabeleceram como uma solução para problemas relacionados com análises espaciais

que utilizam um grande volume de dados (BigData) onde o conhecimento da extensão do problema está incompleto (Lary *et al.*, 2016; Merghadi *et al.*, 2020).

2.4.5.1. Algoritmos de Clustering

A clusterização é uma ferramenta importante na análise de um conjunto de dados. É um método criado para encontrar estruturas de agrupamento dentro dos dados que são caracterizados pelas suas similaridades (Sinaga & Yang, 2020). Os métodos de clusterização aplicados a documentos, dividem-nos em grupos semelhantes de acordo com o seu conteúdo (Gaikwad *et al.* 2014). Este campo de estudo de machine learning tem crescido exponencialmente dado o elevado número de dados que são produzidos em formato de texto (Dodda & Babu, 2024).

Mais especificamente, a análise de clusters permite organizar os padrões encontrados nos dados representados, medidos por vetores ou pontos num espaço multidimensional na forma de agrupamentos baseados na sua similaridade. Os padrões de uma amostra pertencente a um cluster vão ser mais idênticos a outras mesmas amostras pertencentes a esse cluster do que a outras de um cluster diferente (Jain *et al.* 1999). Por norma a qualidade dos clusters é avaliada de acordo com o grau de proximidade do conteúdo dos documentos dentro do mesmo cluster e a diferenças destes documentos para com os restantes clusters, no entanto a ideia de uma clusterização ideal deve ser reformulada de acordo com a escolha de aplicação do método e varia entre diferentes utilizadores (Hotho *et al.* 2005).

O agrupamento de documentos baseado nas suas especifidades pode utilizar diferentes algoritmos dos quais se inclui o *hierarquical clustering* e *k-means* (Renganathan, 2017). O algoritmo de k-means é um algoritmo iterativo que reparte um conjunto de dados num determinado número de clusters, sendo esse número escolhido pelo utilizador, é um algoritmo descritivo o que significa que utiliza os dados para encontrar padrões e estruturas dentro dos mesmos (Wu *et al.*, 2008). Este algoritmo é por norma o mais conhecido e utilizado em técnicas de clustering (Sinage & Yang, 2020) devido à sua simplicidade de implementação o que torna a sua utilização acessível a diversas pessoas (Jain, 2010); pela sua eficiência computacional, especialmente com grandes conjuntos de dados (Wu *et al.*, 2008) e pela adaptabilidade, uma vez que pode ser facilmente adaptado e modificado para atender a requisitos específicos (Arthur & Vassilvitskii, 2007).

Tendo um conjunto de dados D, com n objetos num espaço euclidiano, os métodos de partição distribuem os objetos D em k clusters, C1,…,Ck, Isto avalia a qualidade da partição de modo a que os objetos dentro de um cluster sejam similares entre sim e dissimilares dos objetos em outros clusters. A técnica de partição baseada em centroides utiliza o centroide de um cluster para representar esse cluster, sendo esse centroide definido pela média dos objetos ou pontos atribuídos ao cluster. A diferença entre o objeto e o representante do cluster é definido pela distância euclidiana entre dois pontos x e y. A qualidade desse cluster pode ser medida pela variação intra-cluster, sendo a soma dos erros quadrados entre todos os objetos em Ci e o centroide ci, definida como:

$$E = \sum_{i=1}^{K} \sum_{p \in Ci} dist(p, ci)^2 \quad (2)$$

Onde E é a soma dos erros quadrados para todos os objetos no conjunto de dados; p é o ponto no espaço que representa um objeto; e ci é o centroide do cluster (Han *et al.*, 2011). No entanto encontrar o número apropriado de k (número de clusters) torna-se desafiador pelo facto de se decidir como podemos definir um "bom cluster" (Pham *et al.*, 2005).

Segundo Cui (2020), um dos métodos mais utilizados para definir k é o método do cotovelo (elbow method), sendo aconselhável para valores relativamente pequenos de k. O método do cotovelo calcula a diferença quadrada de diferentes valores de k, quando o k aumenta a média de distorções entre objetos torna-se mais pequena. Sendo que cada objeto se aproxima mais de um centroide, enquanto o k aumenta e o efeito de distorção diminui mais encontra-se o valor ideal de k, sendo que são realizados cálculos para os diversos valores atribuídos a k.

Deste modo, métodos de clusterização podem ser utilizados em diversas áreas, destacando-se: análise de padrões; agrupamento; apoio à decisão e aplicação de machine learning incluindo data mining e análise de documentos; classificação de padrões, etc. (Jain *et al.* 1999). Contudo, um dos maiores desafios presentes na aplicação da clusterização de documentos, sendo esta um campo de Machine Learning e NLP, relaciona-se com a qualidade dos dados, uma vez que esta técnica se baseia na semelhança de frases ou do conteúdo do documento, palavras insignificantes minimizam a precisão dos resultados (Dodda & Babu, 2024).

Capítulo 3 - Dados e métodos

3.1. Enquadramento da área de estudo

Para a presente dissertação foi escolhida como área de estudo as regiões agrárias Entre Douro e Minho; Trás-os-Montes; Beira Litoral; e Beira Interior totalizando 163 municípios pertencentes a 11 distritos do território português (Castelo Branco, Bragança, Guarda, Viseu, Coimbra, Vila Real, Aveiro, Braga, Porto, Viana do Castelo, Leiria dos quais se incluem (Alcobaça, Alvaiázere, Ansião, Batalha, Castanheira de Pêra, Figueiró dos Vinhos, Leiria, Marinha Grande, Pedrógão Grande, Pombal, Porto de Mós) e Santarém, neste último distrito apenas o município de Mação foi incluindo na análise), totalizando aproximadamente 46 000 km^2 de extensão, praticamente metade do território de Portugal Continental (Figura 3). A escolha da área de estudo está relacionada com o projeto de investigação AGROECODECIPHER, visto que estes os municípios tinham a informação tratada.

De notar que o no ano de realização destes inquéritos os municípios da Trofa e Vizela ainda não existiam, tendo sindo anexados aos de Santo Tirso e Guimarães, respetivamente.

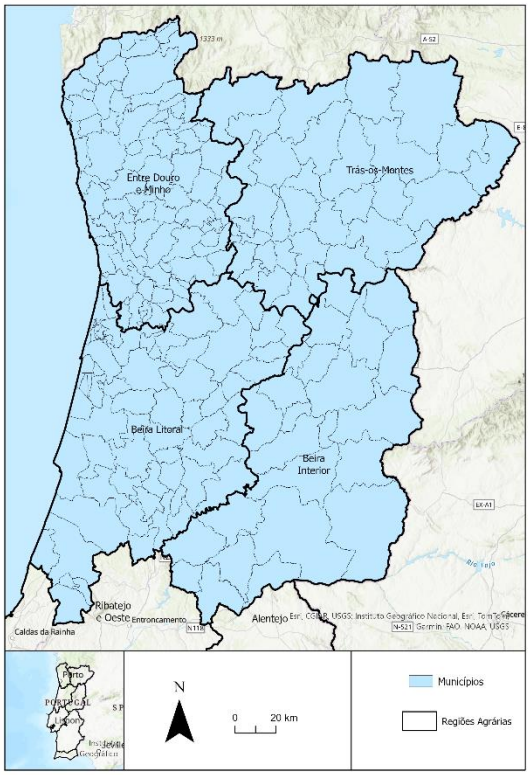


Figura 3 – Área de estudo (Beira Alta, Beira Litoral, Trás-os-Montes e Entre Douro e Minho)

No início dos anos 1950, Portugal era um país essencialmente agrícola, onde este setor representava cerca de 30% do PIB português e a população ativa do setor representava cerca de 40% da população ativa total. Era uma agricultura maioritariamente baseada em métodos tradicionais, pouco evoluídos e mal remunerada para os seus trabalhadores (Avillez, 2016). De acordo com as estimativas do INE os distritos que perfazem a área de estudo totalizavam cerca de 57% da população portuguesa com 4 882 208 habitantes no ano de 1950. Contudo, verificava-se alguma disparidade na distribuição da população, onde Porto, Braga e Aveiro se destacam dos restantes distritos com 533, 226 e 199 mil valores populacionais, respetivamente.

3.2. Dados

A fonte primária de informação desta dissertação são 163 Inquéritos Agrícolas e Florestais do Plano do Fomento Agrário da década de 1950. Cada inquérito contém informações textuais, tabulares e cartografia, recolhida ao longo de alguns anos da década

de 1950, relacionados com as características edafoclimáticas, agroecológicas e fisiográficas (como clima, solo, água, uso do solo e redes rodoviárias) de cada município em estudo. A informação presente nos inquéritos aborda praticamente todos os aspetos da economia rural de Portugal, incluindo uma descrição minuciosa dos usos agrícolas e florestais e sua adequação. Além dos textos os inquéritos também incluem cartografia das condições físicas do território, bem como divisões agrícolas e sub-agrícolas.

Estes inquéritos foram elaborados por engenheiros agrónomos e silvicultores que destacam as particularidades agrícolas e florestais de cada município e estão divididos em três capítulos principais:

1. Inquérito Agronómico: descreve a geografia e topografia do município incluindo geologia, agrologia e recursos hídricos, sistema de transporte e comunicações necessárias para o desenvolvimento agrícola. É também apresentada uma descrição das culturas existentes, técnicas de cultivo e práticas e rotação de culturas, e também é abordado o uso do estrume e adubo bem como os métodos utlizados para o controlo de doenças e pragas. De igual forma é feita a descrição das indústrias agrícolas como a vinícola, oleícola. Na fase final deste capítulo é feita uma análise à produção e consumo de produtos agrícolas, identificando produtos insuficientes e excedentes. São ainda descritos os pormenores da população ativa do setor sendo e o tipo de infraestruturas utilizadas para o armazenamento e alojamento de animais;
2. Inquérito Florestal: é feita a descrição dos maciços florestal e a importância económica e social da silvicultura, propriedades florestais e as técnicas utlizadas para as mesmas. Ainda neste contexto, é descrita a transformação de terrenos incultos e baldios bem como a fixação dos terrenos erodidos;
3. Problemas do município: expõe as principais debilidades agrícolas e florestais no município.

Nesta dissertação apenas será utilizado o Capítulo I – Inquérito Agronómico.

3.3. Metodologia

A Figura 4 apresenta o esquema metodológico que se divide em duas etapas principais com várias tarefas associadas: (a) recolha e tratamento dos dados incluindo

digitalização dos inquéritos agrícolas, utilização de OCR para a passagem do texto físico para texto editável e digital, assim como normalização e limpeza dos dados; (b) Análise espacial do conteúdo presente nos inquéritos com recurso a diferentes métodos.

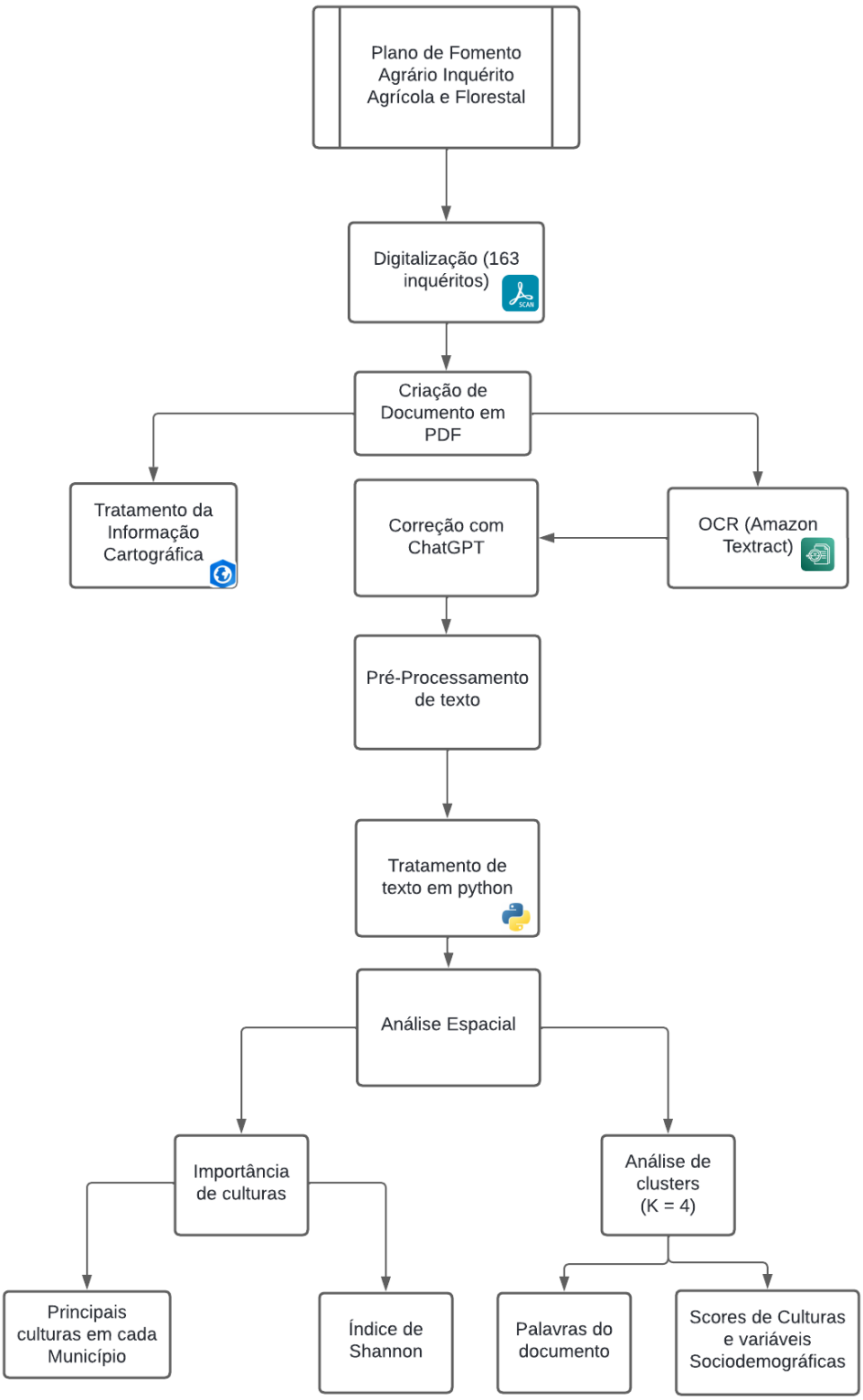


Figura 4 - Fluxograma do procedimento metodológico

3.3.1. Digitalização

A primeira tarefa associada à primeira etapa metodológica consistiu na digitalização de cada página dos 163 inquéritos correspondentes à área de estudo. O objetivo desta tarefa passa por converter a informação em formato físico (analógico/papel) para o digital. Para conseguir efetuar este procedimento foi utilizada a aplicação AdobeScan, disponível para download em qualquer loja de aplicações digitais para *smartphones*. Uma das vantagens encontradas no uso do AdobeScan software foi a sua capacidade para detetar automaticamente os cantos de cada página. Outra vantagem importante foi a possibilidade de garantir uma boa qualidade de digitalização, com bom contraste entre pretos e brancos na imagem (Holley *et. al,* 2009) através da aplicação de um filtro a cada página que permitiu efetuar a binarização das cores para preto e branco. Posteriormente, todos os inquéritos foram armazenados individualmente em formato PDF.

3.3.2. Tratamento da informação cartográfica e vectorização

Após a digitalização de todos os inquéritos separaram-se as páginas que continham informação cartográfica, desde a divisão com as culturas plantadas, a características físicas como serras e rios. Estas páginas foram armazenadas em formato JPG. A informação cartográfica foi posteriormente tratada através de georreferenciação e vectorização. Nesta dissertação, será apresentado apenas o resultado da vetorização dos mapas agrícolas intitulados "Zonas Agrárias". Contudo, destaca-se que nem todos os municípios analisados (n=163) possuíam mapas agrícolas nos seus inquéritos

Assim, numa primeira fase procedeu-se à georreferenciação dos mapas com recurso ao software ArcGISPro. Uma vez que esta informação está disponibilizada ao nível do município, recorreu-se à camada de dados CAOP com as divisões administrativas de Portugal Continental, onde foi criado um script em Python para exportar cada um dos municípios de forma separada. Os limites foram sobrepostos com a ferramenta de adição de pontos de controlo, utilizando no mínimo quatro pontos em cada um dos mapas, nos quais foi aplicada uma transformação da imagem com base num polinómio de primeiro grau. Desta forma, os mapas foram sobrepostos coma *shapefile da CAOP ao nível* municipal. Numa segunda fase, foi necessário realizar a divisão da mesma com recurso à

ferramenta de edição *split,* onde foi vetorizada apenas a linha de separação das divisões agrícolas com identificação das diferentes categorias agrícolas. Devido à indisponibilidade de tempo, não foi possível calcular o EMQ da georreferenciação realizada para os mapas dos documentos.

3.3.3. Utilização do OCR

Uma vez digitalizados os inquéritos, não foi possível utilizar o texto diretamente, pois o sistema não reconhecia os caracteres. Por isso, foi necessário recorrer à tecnologia OCR, com o objetivo de tornar os caracteres reconhecíveis pelo sistema. Para identificar os melhores softwares de OCR, considerando o contexto atual da inteligência artificial generativa, utilizou-se o ChatGPT, na versão 3.5, solicitando que indicasse os principais softwares OCR para auxiliar no reconhecimento de caracteres. A resposta incluí cinco softwares distintos: Adobe Acrobat, Microsoft OneNote, Abbyy FineReader, Tesseract e Amazon Textract. Especificamente, os softwares Adobe Acrobat e Microsoft OneNote não têm como foco principal ferramentas de OCR, e são ambos pagos e não open-source. Já o software Abbyy FineReader foi desenvolvido com especial foco na tecnologia OCR, convertendo documentos digitais e PDFs em texto editável; é pago e não open-source. O Tesseract, criado e mantido pela Google, utiliza a mesma tecnologia, sendo gratuito e open-source. O Amazon Textract é considerado um dos melhores softwares disponíveis atualmente; e é um serviço pago baseado em cloud e não open-source. Por último, e apesar de não estar incluído na resposta do ChatGPT, também considerámos o software Transkribus que surgiu em contexto académico, focando-se essencialmente na transformação de documentos históricos em texto editável. Apresenta um plano gratuito e outro pago, e, como a maioria destes softwares, também não é open-source.

Para a escolha final do software OCR a ser utilizado nesta dissertação foi realizado um teste de precisão entre as seis hipóteses descritas anteriormente. Com base numa página modelo do inquérito, que continha 202 caracteres, esta foi introduzida em cada um dos seis softwares e verificou-se o nível de precisão com base na comparação do resultado do software e da folha modelo original, por exemplo, se o inquérito original tinha 202 caracteres e no ficheiro final do OCR desses 202 apenas 90 estavam corretos, através de uma regra de 3 simples obtém-se uma precisão de 44%. Verificou-se que dos softwares testados os que apresentaram os melhores resultados de precisão foram a Amazon

Textract e Transkribus (Quadro 1). A escolha final do software recaiu para aquele que possuiu a percentagem de precisão mais elevada e custo de utilização mais reduzido: Amazon Textract. Deste modo, utilizou-se o software Amazon Textract para reconhecimento textual dos 163 inquéritos. Os resultados do OCR foram posteriormente armazenados em formato txt.

Adobe Acrobat	12,4%
Microsoft OneNote	16,3%
Abby FineReer	51,5%
Tesseract	50,0%
Amazon Textract	72,3%
Transkribus	70,0%

Quadro 1 - Valor de precisão obtido para cada software OCR.

3.3.4. Pós-Processamento dos resultados OCR

Embora úteis, as ferramentas de OCR não são suficientes para garantir um reconhecimento completo do texto presente em cada página dos inquéritos, como também foi observado no teste de precisão realizado anteriormente entre os softwares disponíveis. Esta limitação é ainda mais evidente em inquéritos históricos com mais de 70 anos, nos quais algumas páginas apresentam texto pouco percetível devido ao desbotamento da tinta ou a manchas causadas pela sua antiguidade. No Quadro 2, apresentam-se os valores de precisão do software OCR Amazon Textract para a primeira página de 19 inquéritos escolhidos aleatoriamente. Observa-se que de todos os inquéritos apenas cincos apresentaram valores de precisão superiores a 80%. Com um valor de precisão médio de 58,41%, com o inquérito do município de Vinhais a apresentar o valor mais baixo (18,90%) e o inquérito do município de Cinfães a apresentar o valor mais alto (97,55%), com uma margem de erro médio de 41,59%. Como se pode observar, embora alguns inquéritos apresentem um elevado valor de precisão, muitos exibem uma margem de erro de aproximadamente 28%. Este valor pode indicar que o inquérito não contém toda a informação textual original, que a formatação do texto pode estar incorreta, ou que existem erros ortográficos.

Município	Nº de caracteres (folha original)	Nº de caracteres certos depois do OCR	Precisão OCR
Armamar	220	130	59,09%
Arouca	148	80	54,05%
Boticas	156	87	55,77%
Caminha	134	45	33,58%
Castelo de Paiva	112	60	53,57%
Cinfães	163	159	97,55%
Góis	203	196	96,55%
Ílhavo	212	197	92,92%
Mesão Frio	135	39	28,89%
Mirandela	206	194	94,17%
Mortágua	136	109	80,15%
Paredes	220	85	38,64%
São Pedro do Sul	134	29	21,64%
Vagos	118	60	50,85%
Vila Real	144	53	36,81%
Vinhais	164	31	18,90%
Viseu	204	135	66,18%
Esposende	164	112	68,29%
Terras de Bouro	119	74	62,18%

Quadro 2 - Valor de precisão OCR para a primeira página de 19 inquéritos escolhidos aleatoriamente

Como tal, foi necessário desenvolver uma estratégia para garantir a integridade total da informação presente em cada inquérito. Optou-se por recorrer à inteligência artificial generativa através do ChatGPT que permite desempenhar tarefas de processamento de linguagem natural, incluindo a correção de texto (Eke, 2023). Assim, a fase final do tratamento dos inquéritos consistiu em colocar o resultado OCR de cada inquérito no ChatGPT na versão 4 com o objetivo de corrigir o texto na integra, i.e., garantir que o texto final seja o mais semelhante/igual ao original. Contudo, destaca-se que podem ainda existir algumas diferenças (e.g., palavras, formatação) entre o inquérito original e o resultado. A utilização do ChatGPT para correção do texto permitiu aumentar, para todos os inquéritos, o valor de precisão acima dos 90% (Quadro 3). Destaca-se o inquérito do município Armarmar com o menor valor (92,73%). A precisão média utilizando o ChatGPT foi de 97.42%, incrementando em média 39,01% o valor de precisão dos resultados do OCR.

Município	Nº de caracteres (folha original)	Nº de caracteres certos depois do ChatGPT	Precisão ChatGPT	Incremento de precisão
Armamar	220	204	92,73%	33,64%
Arouca	148	145	97,97%	43,92%
Boticas	156	156	100,00%	44,23%
Caminha	134	126	94,03%	60,45%
Castelo de Paiva	112	111	99,11%	45,54%
Cinfães	163	163	100,00%	2,45%
Góis	203	203	100,00%	3,45%
Ílhavo	212	212	100,00%	7,08%
Mesão Frio	135	133	98,52%	69,63%
Mirandela	206	206	100,00%	5,83%
Mortágua	136	132	97,06%	16,91%
Paredes	220	214	97,27%	58,64%
São Pedro do Sul	134	127	94,78%	73,13%
Vagos	118	115	97,46%	46,61%
Vila Real	144	138	95,83%	59,03%
Vinhais	164	158	96,34%	77,44%
Viseu	204	194	95,10%	28,92%
Esposende	164	161	98,17%	29,88%
Terras de Bouro	119	115	96,64%	34,45%

Quadro 3 - Valor de precisão após correção com o ChatGPT para a primeira página de 19 inquéritos escolhidos aleatoriamente

Após a correção final foi necessário converter os resultados para formato txt e selecionar o texto do capítulo I - Inquérito Agronómico. Para tal, foi utilizada a linguagem de programação Python. A escolha do Python baseou-se na sua natureza dinâmica e na sua otimização para a execução de scripts, bem como na existência de diversas bibliotecas desenvolvidas para seleção e tratamento de texto. Para facilitar a utilização desta linguagem e identificar erros na tipografia, todo o processo em Python foi desenvolvido no editor de código Visual Studio Code.

3.3.5. Análise espacial

3.3.5.1. Matriz de importância das culturas

Para compreender a importância de cada cultura agrícola em cada município, foi necessário, primeiramente, selecionar as culturas a serem analisadas. Para isso, recorreu-se à base de dados do projeto "Agriculture in Portugal- 1870-2010 (FCT-PTDC/HIS-HIS/122589/2010)" (http://www.ruralportugal.ics.ul.pt/data-files/) para identificar as

principais culturas agrícolas. A lista considerada inclui as seguintes culturas: milho, batata, vinha, azeite, arroz, centeio, trigo, nabo, feijão, grão, aveia, oliveira, amêndoa, cevada, chícharo, cortiça, noz, castanha, avelã, tremoço, laranja, limão, cera, mel, sal, vinagre, aguardente, pêssego, pera e maçã. De seguida, em ambiente Python, foram utilizadas as bibliotecas pandas, scikit-learn e TfidfVectorizer para criar um vetor que seria aplicado a cada inquérito, com o objetivo de gerar uma matriz de importância das culturas com base na lista mencionada (scores). Com a biblioteca Tf-idf, foi criado um índice para cada cultura, resultado da vetorização de cada palavra ao longo dos inquéritos. Cada índice foi iterado para todos os inquéritos, permitindo a obtenção da importância de cada cultura, i.e., scores, em cada inquérito. Finalmente, os resultados foram gravados numa matriz através da biblioteca pandas, possibilitando o cálculo da correlação de Pearson entre as culturas.

O coeficiente de correlação de Pearson é uma medida estatística que quantifica a relação entre duas variáveis quantitativas, tendo sido desenvolvido pelo matemático Karl Pearson no início do século XX. O coeficiente varia entre -1 e 1 onde:

- r = 1 indica uma correlação linear positiva perfeita
- r = -1 indica uma correlação linear negativa perfeita
- r = 0 indica que não há correlação

Segundo Cohen (1988) valores entre 0,1 e 0,3 indicam uma correlação fraca; valores entre 0,3 e 0,5 indicam uma correlação modera e valores acima de 0,5 indicam uma correlação forte. Para esta dissertação o coeficiente de Pearson foi calculado com auxílio da biblioteca pandas, tendo sido necessário converter a tabela de scores numa data frame à qual foi aplicada a função .corr tendo sido escolhido o método ‘pearson’ para este cálculo,

Nesta fase, foi também calculado o Índice de Shannon para avaliar a diversidade das culturas em cada município. Este índice permite identificar os municípios que praticam um cultivo mais diversificado e aqueles que se concentram em culturas específicas. O cálculo do Índice de Shannon foi realizado no software Excel, utilizando a matriz obtida anteriormente, onde cada linha representa um município. Primeiro, foi calculado o valor total das culturas para cada linha. Em seguida, foi determinada a proporção de cada cultura, dividindo o valor de cada cultura pelo total da linha. Depois, multiplicou-se a proporção de cada cultura pelo logaritmo natural dessa proporção. Os resultados dessas multiplicações foram somados para cada cultura. Por fim, para obter o

Índice de Shannon de cada município, multiplicou-se a soma dos valores de cada cultura por -1.

$$H = -\sum_{I=1}^{S} Pi\ lnPi$$

3.3.5.2. Análise de Clusters

Antes de proceder com a análise ao conteúdo dos inquéritos, foi efetuado um pré-processamento comum a todos eles. Nesta etapa, utilizou-se a biblioteca NLTK, associada ao processamento de linguagem natural, para descarregar um conjunto de "stop words" - palavras frequentemente utilizadas, como artigos definidos, advérbios, entre outros, que não acrescentam informação relevante para o estudo. Esta prática é comum na aplicação de NLP e *text mining*. Posteriormente, foi necessário criar uma lista personalizada com mais *stop words* para melhorar os resultados e refinar o algoritmo, em baixo segue a lista de algumas *stops words* utilizadas (Anexo 1):

'kg/ha', 'principalmente', '-', '|', 'cerca', ',', '.', 'especialmente', 'Quadro', 'geralmente', 'se', 'onde', 'grande', 'kg', 'relativamente', 'sendo', 'alguma', 'maior', 'pode', 'embora', 'parte', 'quase', 'devido', 'e', 'e,', 'sempre', 'geral,', 'geral', 'devido', '1', 'casos,', 'casos', 'neste', 'regra,', 'regra', 'Quadro', 'entanto,', 'forma', 'são,', 'quantidade', 'via',

Outro procedimento realizado, a normalização da tipografia, assegurando uniformidade durante as aplicações e análises do código sobre o conteúdo textual. Por fim, procedeu-se à vetorização do texto utilizando o TfidfVectorizer, uma função da biblioteca scikit-learn. Esta função serve para criar uma matriz numérica que tem em conta a frequência de termos (TF). Optou-se por utilizar a sua versão inversa, pois desta forma é possível encontrar termos que são frequentes e raros ao mesmo tempo. A sua raridade prende-se com a frequência de um certo termo numa parte do texto que deixa de ser frequente noutra.

Após o pré-processamento, utilizou-se a biblioteca scikit-learn para proceder com a análise de clusters. A análise de clusters é utilizada como uma ferramenta para compreender as relações entre o conteúdo dos 163 inquéritos agrícolas ao nível do município por forma a permitir a identificação de padrões. Esta técnica agrupa os municípios com base nas características do seu conteúdo, considerando o número de ocorrências e a localização das palavras nos inquéritos. Dado que no conjunto dos 163

inquéritos se totalizam mais de 20.000 páginas, torna-se impraticável realizar este tipo de análise manualmente. Assim, a análise de clusters identifica tendências ocultas nos dados, formulando a hipótese de agrupamento mais fiável através do algoritmo de clusterização K-means. Esta abordagem permitirá avaliar como os municípios se agrupam com base nos conteúdos dos inquéritos agrícolas. Primeiramente foi necessário definir o número máximo de clusters (valor de k), sendo que este valor deve ser baseado no "*elbow method*". Foram feitas tentativas automáticas com diferentes valores de k no intervalo de [1, 11], e foi criado um gráfico de linha para cada um desses testes. O ponto que cria uma concavidade no gráfico corresponde ao número indicado de clusters. A análise da Figura 5 indicou que o valor ideal de k é 4, uma vez que é o último valor relativamente constante antes de uma queda acentuada no desempenho. Embora pudesse surgir a dúvida entre escolher k igual a 3 ou 4, optou-se pelo valor mais elevado para obter uma segmentação mais detalhada, resultando em um maior número de clusters. Assim, foram criados quatro clusters diferentes, os quais foram visualizados através de um gráfico criado com a biblioteca matplotlib. Posteriormente, recorreu-se à biblioteca pandas para criar um *dataframe* com as palavras mais recorrentes de cada município e o respetivo número de vezes que se repetiam. Desta forma, foi possível visualizar quais os termos que mais se repetiam e que explicavam a divisão dos clusters para entender as especificidades agrícolas regionais.

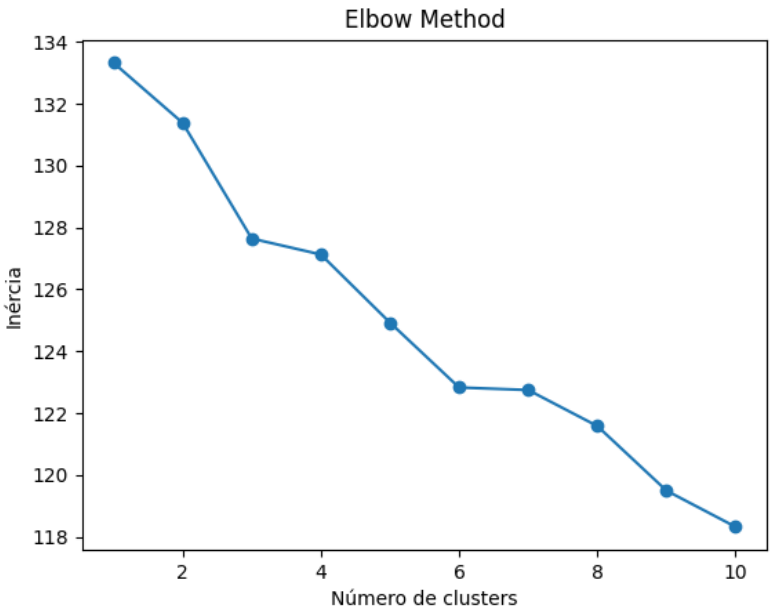


Figura 5 - Desempenho de k ao longo do intervalo de valores (Método do Cotovelo)

Por fim, procurou-se verificar se os clusters obtidos poderiam ter alguma relação com determinadas características da população agrícola ou o tipo de cultivo da época. Para tal, foi elaborada uma tabela com informações sobre a importância de cada cultura, juntamente com variáveis demográficas extraídas do IX Recenseamento Geral da População (1950) do INE, especificamente do Tomo III, volumes 1 e 2. As variáveis incluíram a população agrícola masculina e feminina, com subdivisões em categorias como idades entre 12 e 14 anos, população ativa e indivíduos com mais de 65 anos. Esta tabela foi utilizada para, utilizando o algoritmo k-means, identificar os clusters formados por esse novo conjunto de dados, aplicando o mesmo valor de k, de forma que os resultados sejam comparáveis com o mapa de clusters gerado anteriormente (Wall ,1986)

Capítulo 4 - Análise e discussão de Resultados

4.1 Vectorização dos mapas agrícolas

A Figura 6 apresenta o resultado do tratamento da informação cartográfica existente nos inquéritos. Como se observa na legenda da Figura 6, durante a execução destes inquéritos, não havia diretrizes claras para a elaboração da nomenclatura, o que resultou em legendas pouco informativas e discrepâncias na escolha das nomenclaturas para as diferentes categorias, com cerca de 135 classes distintas. Esta situação complicou a representação cartográfica, visto que, frequentemente, a mesma classe de uso e ocupação do solo era representada por nomenclaturas diferentes, dificultando a distinção semântica das classes no mapa final. Com o intuito de facilitar a visualização e compreensão do mapa, procedeu-se a uma proposta de agrupamento das diferentes classes em 10 categorias, conforme apresentado na Figura 7. Verificou-se que a leitura das classes se tornou mais clara; no entanto, dado o título "Zonas Agrárias", seria expectável encontrar mais informações relacionadas com a agricultura, para além disso, importa referir que a primeira classe não possui descrição uma vez que foi um conjunto de classes previamente recolhidas que possuíam apenas informação numérica que nada acrescentam à interpretação do mapa. Contudo, observa-se que grande parte da informação está associada a aspetos físicos do território. A classe "região numerada" refere-se a áreas

delimitadas às quais foi atribuído um número, sendo este associado a uma descrição específica. Importa ainda destacar que cada região pode incluir vários inquéritos, e a descrição pode variar entre eles.

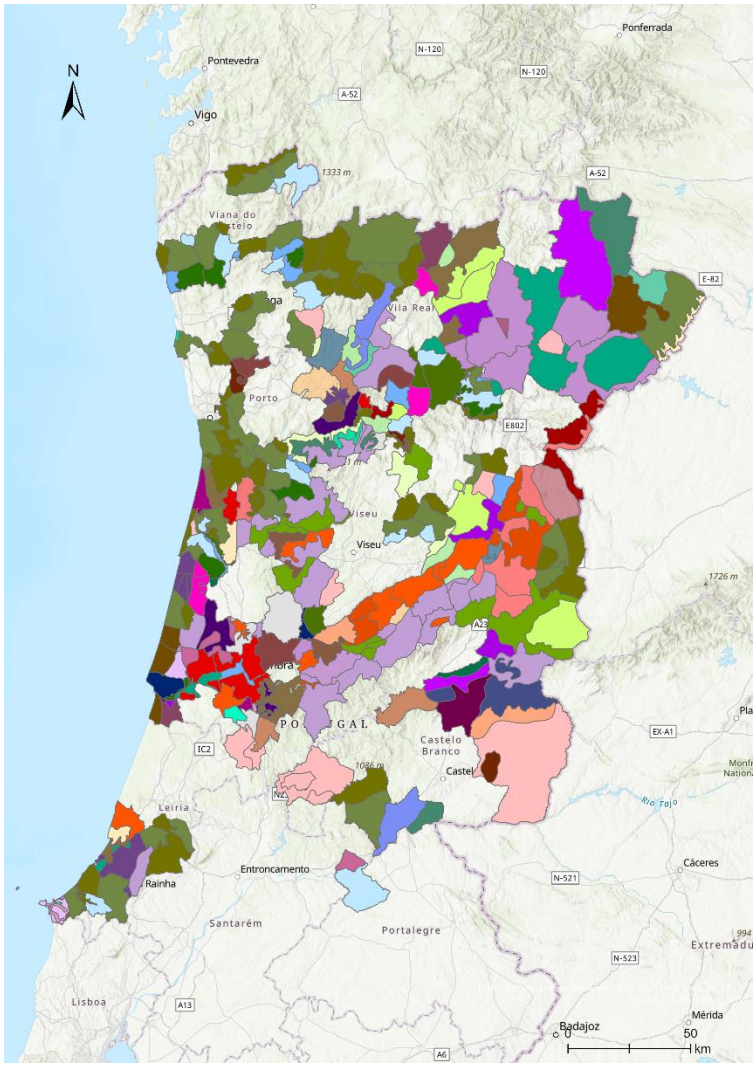


Figura 6 – Zonas agrárias (Inquéritos de 1950)

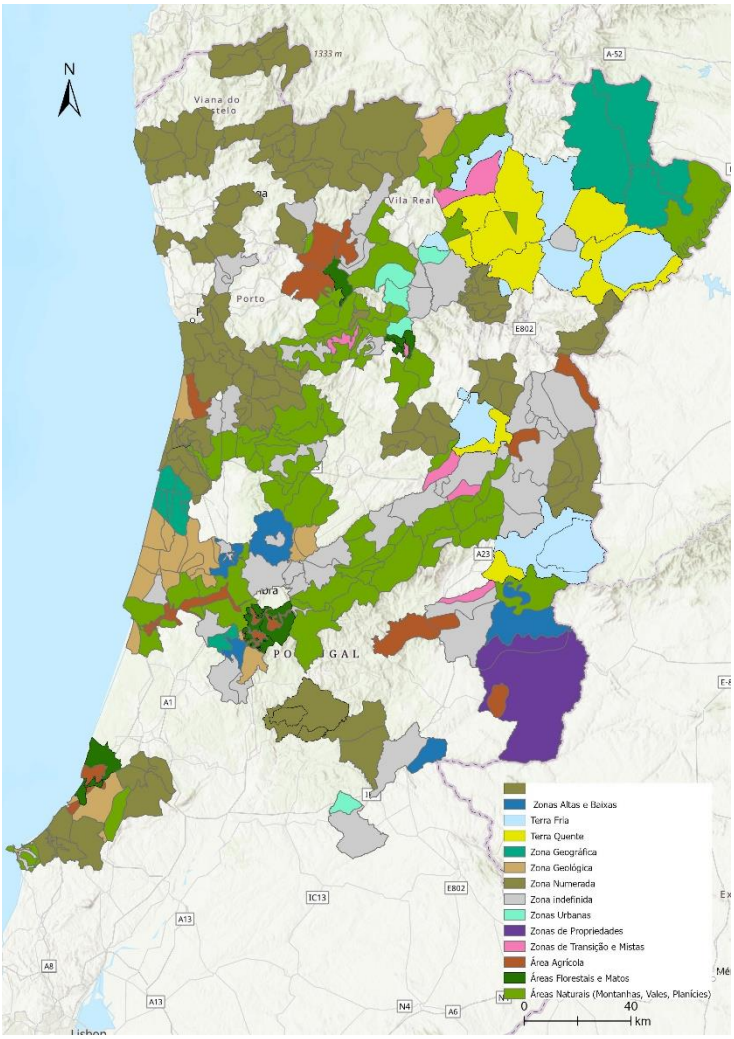


Figura 7 – Zonas agrárias com proposta de agrupamento de classes agrícolas (Inquéritos de 1950)

4.2. Importância das Culturas

Primeiramente foi necessário proceder ao cálculo de uma matriz de importância dos produtos agrícolas baseada na frequência e a sua posição relativa no inquérito de cada município, sendo apresentado um excerto desta tabela no Quadro 4 com os primeiros 6 municípios, no entanto foram calculados os valores para todos os 163 municípios que fazem parte do estudo (Anexo 2).

	Inquéritos

		Aguiar Da Beira	Albergaria-a-Velha	Alcobaça	Alfândega	Alijó	Almeida
Culturas	milho	58,1	58,1	50,9	0,0	14,3	0,0
	batata	35,2	47,3	24,1	21,1	14,1	7,0
	vinha	27,5	10,2	28,5	10,2	10,2	10,2
	azeite	10,8	5,4	6,5	5,4	6,5	5,4
	arroz	1,2	28,7	35,8	1,2	2,4	2,4
	centeio	35,6	14,3	6,1	40,7	9,2	12,2
	trigo	13,0	15,0	25,0	27,0	6,0	13,0
	nabo	9,4	0,0	0,0	14,1	0,0	0,0
	feijão	17,4	14,3	17,4	1,0	9,2	2,0
	grão	8,2	2,3	8,2	1,2	2,3	0,0
	aveia	3,7	16,0	25,9	4,9	3,7	0,0
	oliveira	17,0	9,6	9,6	9,6	4,3	1,1
	amêndoa	0,0	0,0	0,0	21,7	0,0	0,0
	cevada	12,6	11,3	18,8	6,3	1,3	0,0
	chícharo	0,0	0,0	3,9	0,0	0,0	0,0
	cortiça	2,0	0,0	0,0	0,0	0,0	2,0
	noz	0,0	0,0	0,0	0,0	0,0	0,0
	castanha	7,7	0,0	0,0	0,0	0,0	0,0
	avelã	0,0	0,0	0,0	0,0	0,0	0,0
	tremoço	0,0	7,4	1,5	1,5	0,0	0,0
	laranja	0,0	0,0	8,7	0,0	2,2	0,0
	limão	0,0	0,0	4,2	0,0	4,2	0,0
	mel	2,5	5,0	12,4	1,2	0,0	1,2
	aguardente	0,0	2,8	11,3	1,4	0,0	1,4
	pêssego	0,0	0,0	24,4	0,0	0,0	0,0
	pêra	0,0	0,0	16,0	0,0	0,0	0,0
	maçã	7,0	0,0	16,4	0,0	0,0	2,3

Quadro 4 - Scores Culturas para seis municípios em estudo

A análise da matriz de correlação das diferentes culturas agrícolas, representada na Figura 8, revela as relações entre os plantios estudados, fundamental para compreender como o cultivo de determinadas culturas pode influenciar a produtividade de outras quando plantadas em proximidade ou em rotação. A matriz ilustra o valor e a direção das correlações, onde valores próximos de 1 indicam uma correlação fortemente positiva, enquanto valores próximos de -1 indicam uma correlação fortemente negativa.

Verifica-se que algumas culturas apresentam correlações positivas, como nos conjuntos milho-feijão (0,85), milho-batata (0,63), cevada-aveia (0,56), batata-feijão (0,56) e entre as árvores frutíferas pêssego e maçã (0,51), o que pode sugerir que estas culturas podem coexistir de forma benéfica nos ecossistemas agrícolas, facilitando uma

gestão eficaz do solo e dos recursos utilizados para estes plantios. A correlação alta entre o milho e o feijão pode ser explicada pela forma como as duas competem por fatores como radiação solar, água e nutrientes. Num estudo feito por Portes (1984) verificou-se que a resposta do milho à alta intensidade de luz solar é de crescimento continuo não existindo saturação fotossintética, já o feijão ficou saturado por volta de um terço da luminosidade mais alta aplicada ao milho, justificando que o feijão se consegue desenvolver na sombra do milho. No entanto, o principal fator que beneficia a plantação destas culturas, em conjunto, é o facto da presença do milho amenizar a temperatura mantendo a humidade do solo que o feijão necessita por mais tempo

Por outro lado, a correlação entre o milho e a cortiça (-0,28) e entre o milho e a amêndoa (-0,26), embora sejam correlações moderadamente baixas, podem indicar que o cultivo conjunto destas culturas pode ser desvantajoso. Destaca-se ainda a correlação entre a amêndoa e a castanha, que, apesar de serem ambos frutos secos, apresentam uma correlação baixa, sugerindo que o rendimento destas culturas não está relacionado. Contudo, a razão principal pela qual normalmente não se planta castanheiros (*Castanea sativa*) junto com amendoeiras (*Prunus dulcis*) está relacionada principalmente com as diferenças nas necessidades ecológicas e de gestão de cada espécie. Isto porque os castanheiros são árvores que prosperam em climas mais temperados e húmidos, embora tolerem períodos secos, precisam de maior disponibilidade de água ao contrário da amendoeira que é mais tolerante à seca preferindo verões quentes e secos com chuvas limitadas (Doll *et al*.,2021) (Braga *et al*., 2023).

Apesar de esta análise ser baseada apenas na frequência de palavras e na sua localização no inquérito, sugere algumas relações interessantes para o planeamento agrícola, permitindo otimizar o uso do solo e maximizar o rendimento das culturas. Considerando estas relações entre diferentes cultivos, podem ser tomadas decisões mais informadas sobre a rotação de culturas, promovendo práticas agrícolas mais sustentáveis e eficientes.

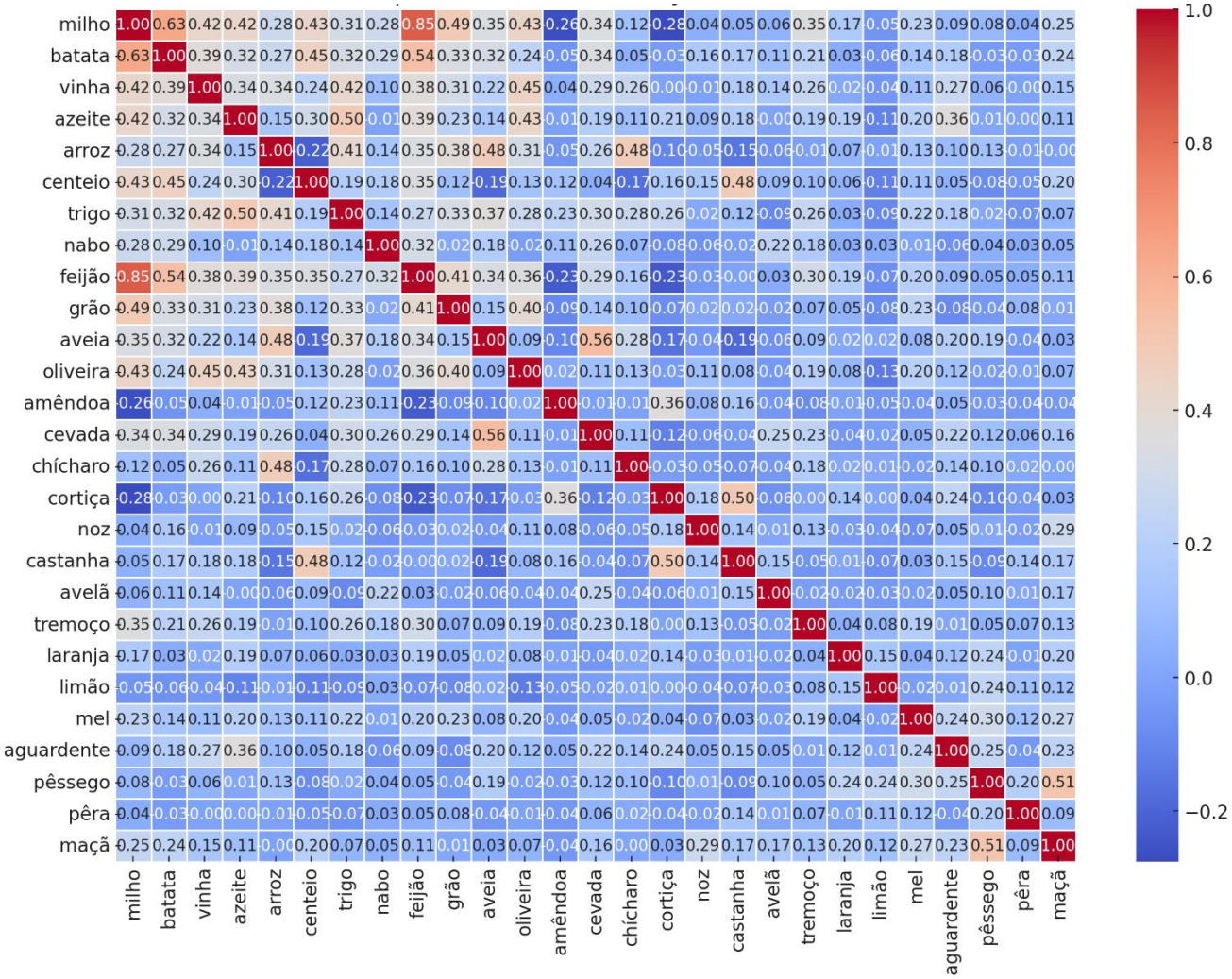


Figura 8 - Matriz de correlação de culturas

O Quadro 5 apresenta o número de municípios em que cada produto agrícola surge como cultura principal. Conforme já mencionado, o milho destaca-se com a maior expressão, sendo a principal cultura em 121 dos 163 municípios analisados, o que corresponde a aproximadamente 74% do total. A batata é a segunda cultura mais predominante, sendo a principal em 12 municípios, representando cerca de 13%. Seguem-se a cortiça, a vinha e o trigo, cada um predominante em 7 municípios, perfazendo aproximadamente 4% do total.

Cultura Principal	Nº de municípios	%
Milho	121	74,23%
Cortiça	7	4,29%
Centeio	4	2,45%
Trigo	7	4,29%
Batata	12	7,36%
Arroz	3	1,84%
Grão	1	0,61%
Azeite	1	0,61%
Vinha	7	4,29%

Quadro 5 – Cultura principal no global dos municípios analisados (%)

A matriz apresentada no Quadro 4 também permitiu identificar a cultura de maior importância, permitindo a elaboração da Figura 9. Como anteriormente referido, o milho é a principal cultura na maioria dos municípios; contudo, observa-se que a região de Trás-os-Montes difere das demais, com uma diversificação dos produtos agrícolas principais, incluindo o centeio, trigo, batata, vinha. A cultura do arroz apresenta pequenas representações, principalmente no litoral do país, mas com pouca expressão. Em contraste, a cultura da batata assume maior relevância, sobretudo na região noroeste do território continental português. A vinha destaca-se particularmente na região de Trás-os-Montes, especialmente nos distritos de Viseu (Tabuaço, São João da Pesqueira, Lamego) e Vila Real (Peso da Régua e Murça), com vários municípios próximos apresentando esta cultura como a de maior importância. No caso do trigo, observa-se uma maior predominância na região interior do país, nomeadamente na Beira Interior e Trás-os-Montes, com maior expressão nos distritos de Bragança (Vinhais e Mirandela) e Castelo Branco (Penamacor e Idanha-a-Nova). Adicionalmente, destaca-se a existência de dois municípios cujas culturas principais diferem das demais: Figueiró dos Vinhos, com a produção de grão, e Proença-a-Nova, com a produção de azeite.

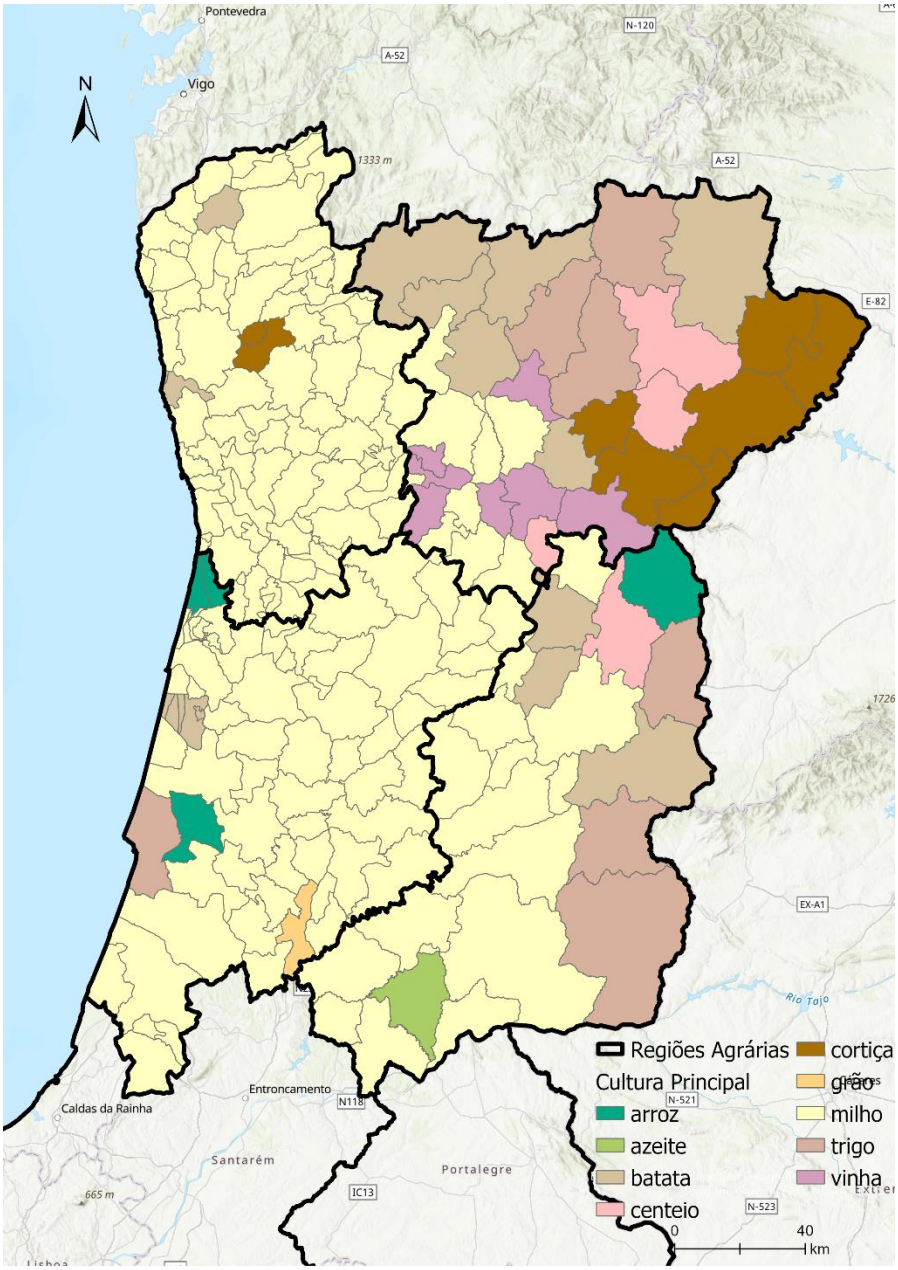
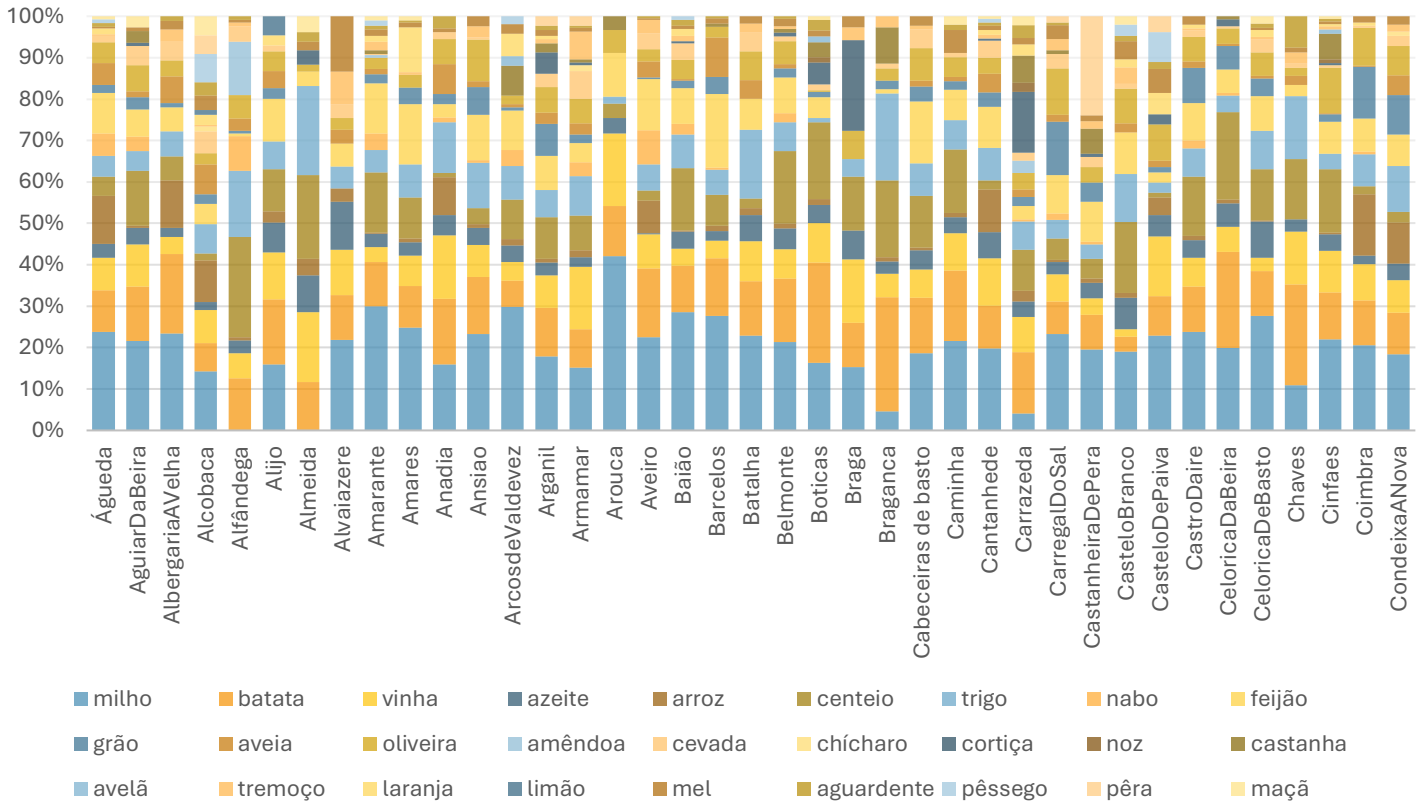
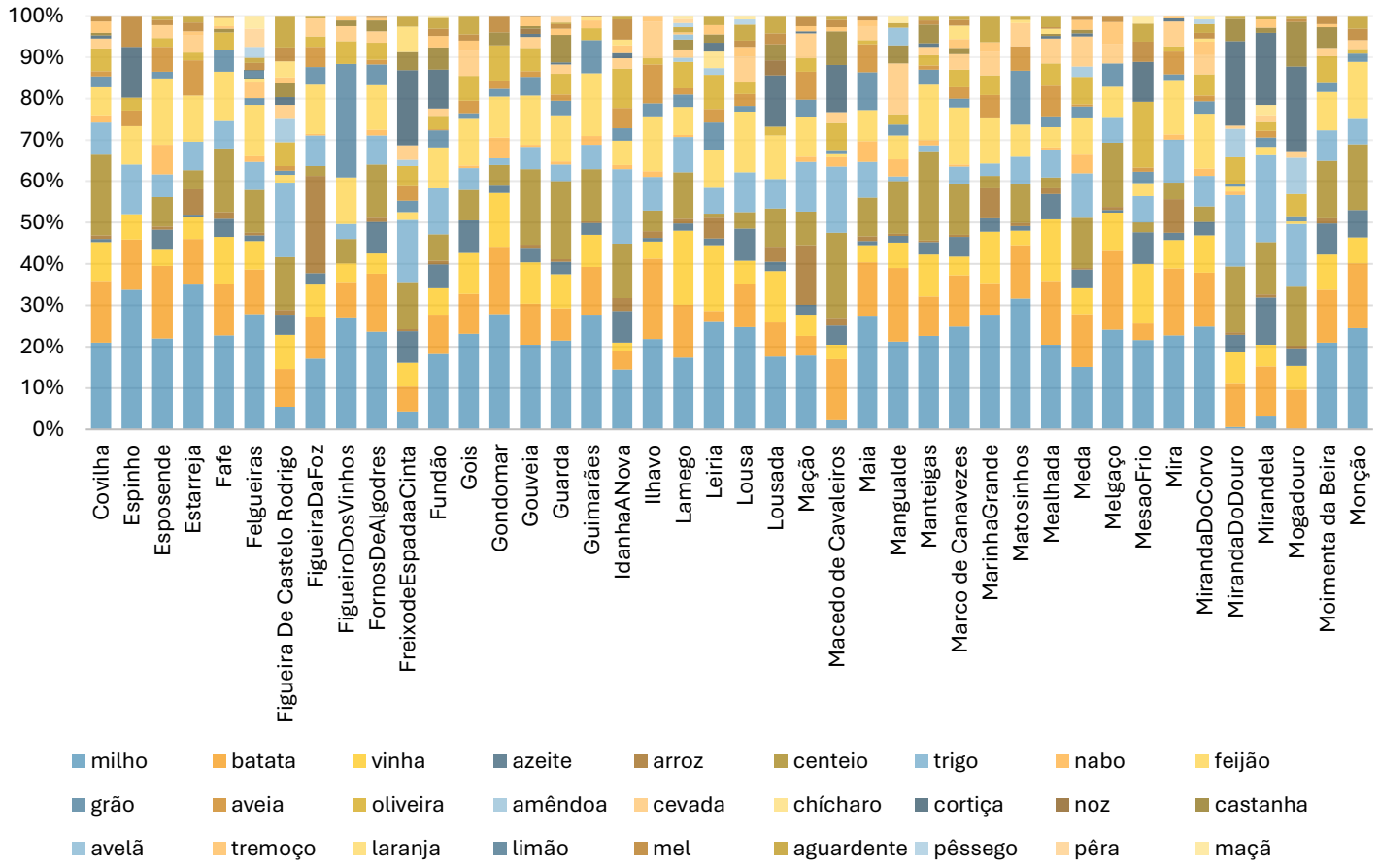


Figura 9 - Cultura Principal por município

De seguida, foi calculada a proporção de cada cultura por município, conforme apresentado na Figura 10. Nesta figura, observa-se que o milho, em conjunto com a batata e a vinha, representa 30% da importância das culturas na maioria dos municípios. As culturas do arroz, azeite e centeio apresentam uma distribuição menos uniforme; no entanto, têm uma importância significativa nos municípios onde estão presentes. Ao analisar os produtos agrícolas individualmente, verificou-se que o milho tem a maior representatividade, com uma média de cerca de 20%, seguido pela batata com 12%, centeio com 10%, e a vinha e o feijão, ambos com 8%. A cultura do feijão destaca-se por

ter uma importância relativamente constante em todos os municípios, podendo ser considerada relativamente elevada. Por outro lado, as árvores de fruto, como o pêssego, laranja, maçã, pera e limão, apresentam menor expressividade devido à baixa proporção de representatividade de produção quando comparadas com os restantes produtos agrícolas.


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Águeda
AguiarDaBeira
AlbergariaAVelha
Alcobaca
Alfândega
Alijo
Almeida
Alvaiazere
Amarante
Amares
Anadia
Ansiao
ArcosdeValdevez
Arganil
Armamar
Arouca
Aveiro
Baião
Barcelos
Batalha
Belmonte
Boticas
Braga
Braganca
Cabeceiras de basto
Caminha
Cantanhede
Carrazeda
CarregalDoSal
CastanheiraDePera
CasteloBranco
CasteloDePaiva
CastroDaire
CeloricaDaBeira
CeloricaDeBasto
Chaves
Cinfaes
Coimbra
CondeixaANova
milho
batata
vinha
azeite
arroz
centeio
trigo
nabo
feijão
grão
aveia
oliveira
amêndoa
cevada
chícharo
cortiça
noz
castanha
avelã
tremoço
laranja
limão
mel
aguardente
pêssego
pêra
maçã


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Covilha
Espinho
Esposende
Estarreja
Fafe
Felgueiras
Figueira De Castelo Rodrigo
FigueiraDaFoz
FigueiroDosVinhos
FornosDeAlgodres
FreixodeEspadaaCinta
Fundão
Gois
Gondomar
Gouveia
Guarda
Guimarães
IdanhaANova
Ilhavo
Lamego
Leiria
Lousa
Lousada
Mação
Macedo de Cavaleiros
Maia
Mangualde
Manteigas
Marco de Canavezes
MarinhaGrande
Matosinhos
Mealhada
Meda
Melgaço
MesaoFrio
Mira
MirandaDoCorvo
MirandaDoDouro
Mirandela
Mogadouro
Moimenta da Beira
Monção
milho
batata
vinha
azeite
arroz
centeio
trigo
nabo
feijão
grão
aveia
oliveira
amêndoa
cevada
chícharo
cortiça
noz
castanha
avelã
tremoço
laranja
limão
mel
aguardente
pêssego
pêra
maçã

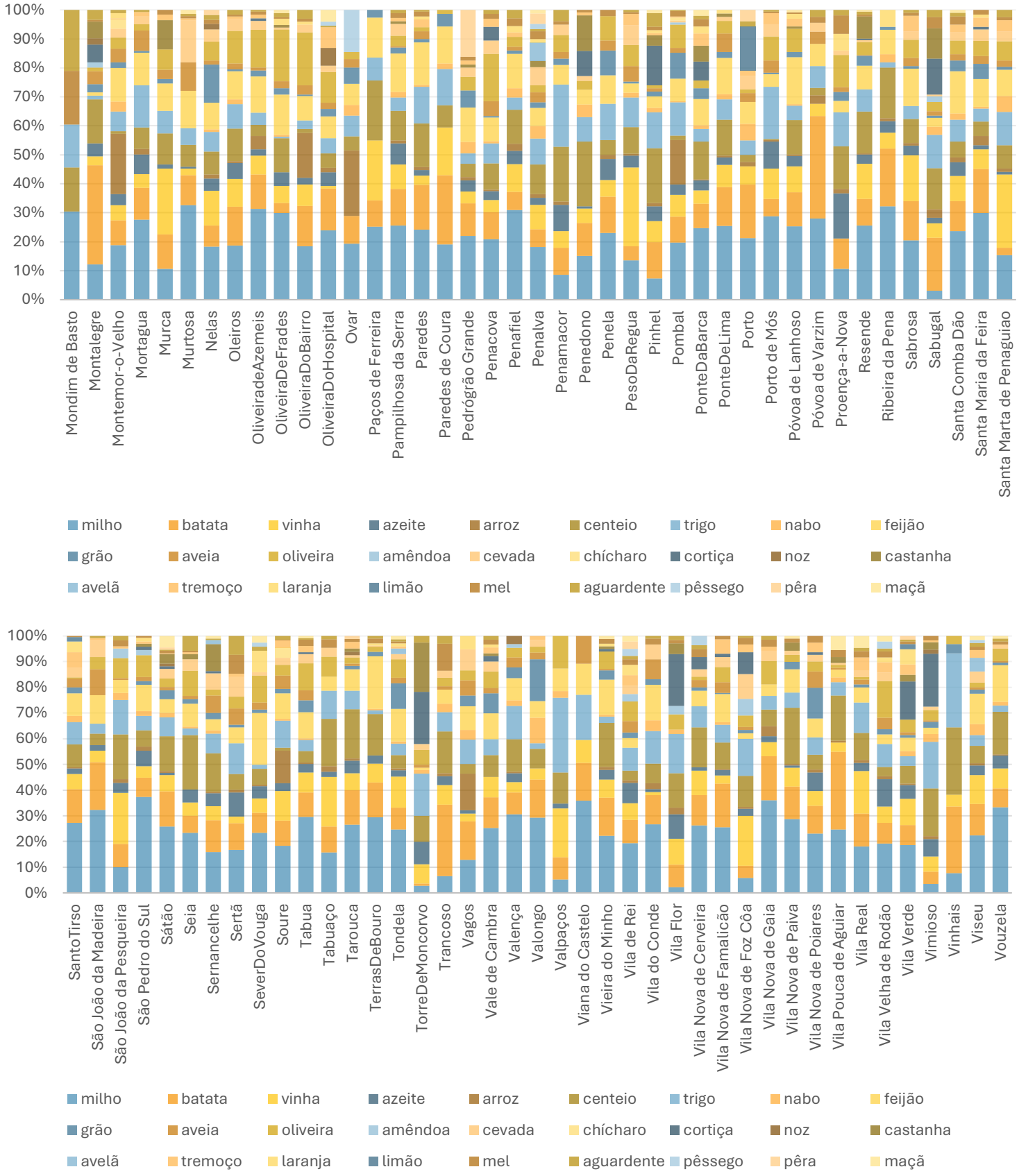


Figura 10 - Proporção de Culturas por município

Como já foi mencionado, o índice de Shannon mede o grau de diversidade de uma área, levando em consideração tanto a riqueza quanto a equitabilidade das espécies. Neste contexto, o índice tem o propósito de aferir a diversidade dos produtos agrícolas produzidos em cada município, verificando a distribuição da sua produção. No contexto desta análise, um valor elevado indica uma maior variedade de produtos agrícolas produzidos, enquanto valores mais baixos sugerem um plantio mais monocultural. A Figura 11 retrata essa distribuição verificando-se que o intervalo de valores mais comum situa-se entre 2 e 2,4. Diferentemente do que foi observado na análise dos produtos agrícolas mais produzidos por município, a região de Trás-os-Montes é aquela com menor diversidade de produtos, registrando os valores mais baixos, com destaque para o município de Mondim de Basto, com aproximadamente 1,5, e Vinhais, com cerca de 1,6 revelando que esta região tende a focar-se mais em certos produtos, não diversificando tanto o seu cultivo.

Em contrapartida, as regiões mais ao sul e ao centro da área de estudo, correspondentes às regiões agrárias da Beira Interior e Beira Litoral, apresentam os maiores índices de diversificação de culturas, com destaque para o município de Alcobaça, com aproximadamente 2,8. Ainda assim, também se observam valores elevados na região de Trás-os-Montes, com Carrazeda de Ansiães apresentando o segundo valor mais elevado de diversificação (≈ 2,7). A região de Entre Douro e Minho possui ainda três municípios com alta taxa de diversificação, sendo eles Felgueiras, Arcos de Valdevez e Vila Verde.

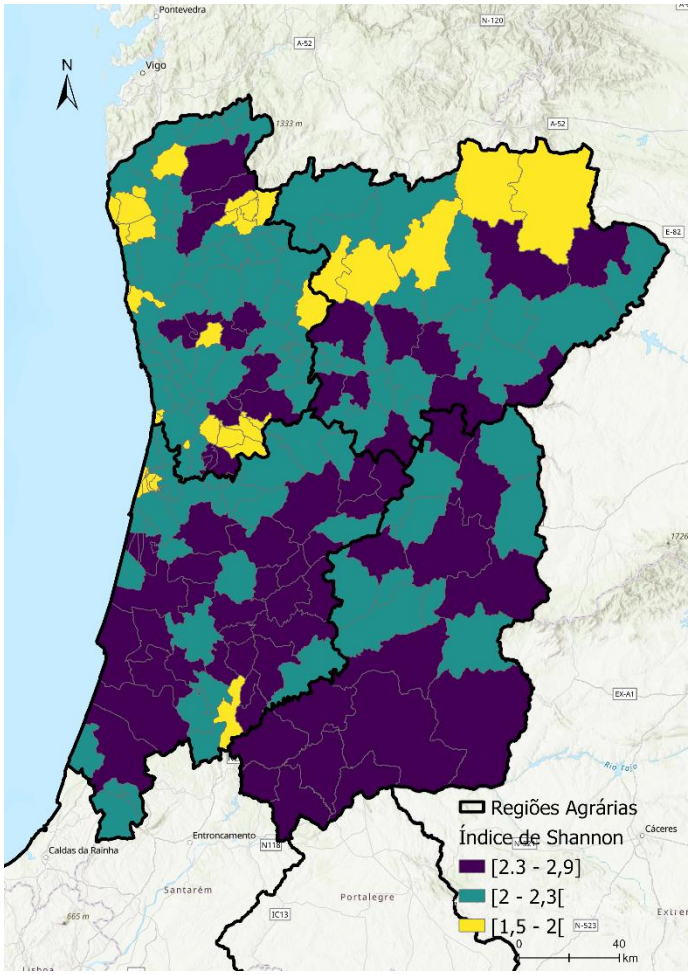


Figura 11 - Índice de Shannon

4.3. Análise de Clusters

Neste estudo, a análise de clusters desempenha um papel crucial, pois permite categorizar cada município das regiões agrárias da Beira Alta, Beira Litoral, Entre Douro e Minho, e Trás-os-Montes com base em características agrícolas específicas. Esta abordagem não só facilita a interpretação dos resultados, mas também permite uma comparação com um conjunto de variáveis explicativas que possam justificar a distribuição dos clusters. Ao agrupar municípios com características agrícolas semelhantes, torna-se possível destacar as principais diferenças entre os grupos, bem como compreender os elementos que os tornam semelhantes.

Para esse propósito, foi elaborado o mapa da Figura 12, que visa compreender a distribuição espacial dos clusters identificados pelo algoritmo, permitindo a visualização dos principais agrupamentos. Observa-se que a distribuição dos clusters por região agrária

e município não é uniforme; existem municípios que, apesar de se situarem numa determinada região agrária e fazerem fronteira com municípios do mesmo distrito/região, apresentam características agrícolas mais próximas às de outra região. A análise do mapa da Figura 12 evidencia vários casos deste fenómeno, sendo o mais notável o do cluster 2, predominantemente concentrado na região Nordeste (Região de Trás-os-Montes), mas que inclui o município de Aguiar da Beira, do distrito da Guarda, localizado na região Centro.

Destaca-se que o cluster 2 é o que possui menor número de municípios, contrastando com o cluster 3, que é o mais representado. A localização do cluster 1 merece atenção especial, dado que se encontra maioritariamente agrupado com municípios de Vila Real e Bragança, sugerindo uma região mais homogénea no seu contexto agrícola, possivelmente explicada por fatores físicos e ambientais comuns a todos os municípios deste cluster.

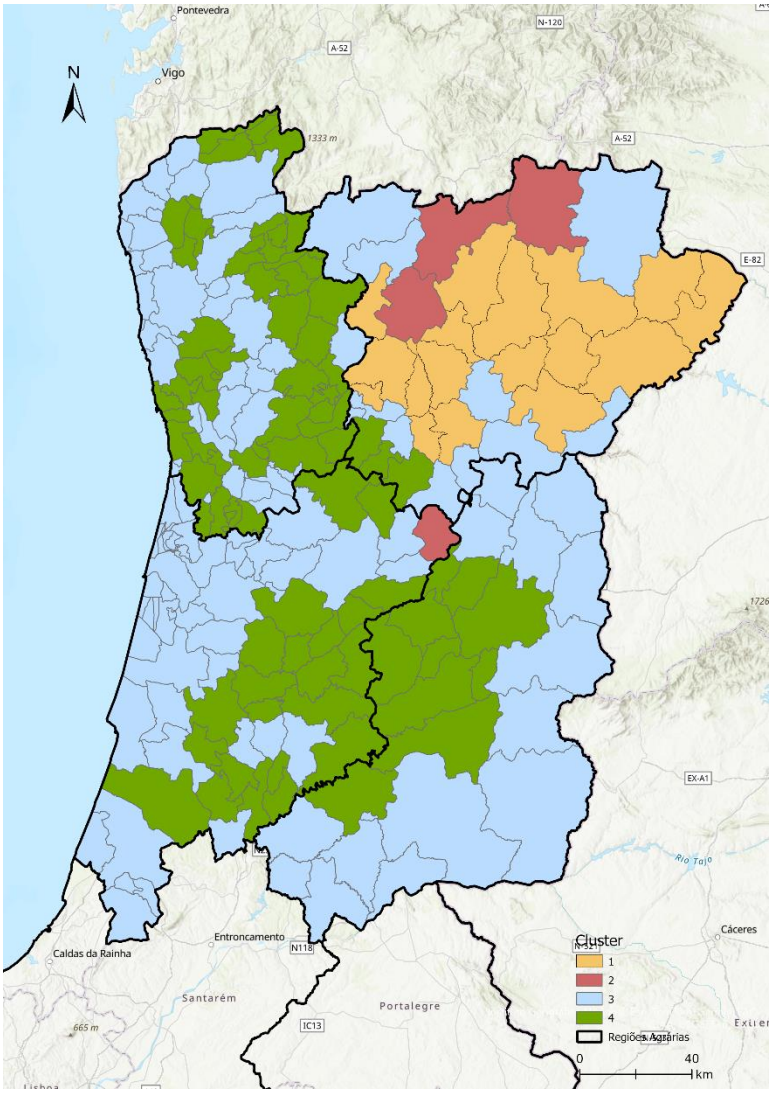


Figura 12 - Distribuição de Clusters (K = 4)

Segue-se uma análise detalhada de cada cluster representado na Figura 12, destacando-se as palavras mais relevantes em cada um, utilizando uma nuvem de palavras. Importa referir que apesar de diversas tentativas não foi possível remover o termo "Se" no entanto, o mesmo não afetou os resultados, sendo apenas um problema visual na nuvem de palavras. Iniciando pelo Cluster 1, anteriormente mencionado, observa-se que os municípios pertencentes estão maioritariamente agrupados na região de Trás-os-Montes. As palavras que mais influenciaram a formação deste cluster, por ordem de relevância, foram: centeio, vinho, trigo, vinha, solo e Douro, conforme ilustrado na Figura 13. Destes termos é especialmente importante destacar "Vinha", "Vinho" e "Douro", e da sua relação com a região do Douro que é famosa pela produção de vinhos de excelente qualidade, pois o rio Douro e os seus afluentes fornecem água para a

irrigação das vinhas. O curso do rio entre os vales ajuda no cultivo da uva uma vez que as protege do vento e servem como um regularizador de temperatura.

Figura 13 - Nuvem de Palavras Recorrentes (Cluster 1)

Ao analisar o Cluster 2, observa-se uma distribuição predominantemente homogénea, com a exceção do município de Aguiar da Beira, localizado no centro da área de estudo. Conforme apresentado na Figura 14, as palavras que mais se destacam e caracterizam este cluster são batata, milho, centeio, agrícolas, exploração e alimentação.

O que diferencia este cluster dos restantes é a presença de termos relacionados com a dependencia da água, surgindo o "sequeiro" revelando que os munícipios do cluster 2 adotam práticas agricolas menos dependentes da rega, estabelecendo-se essencialmente culturas que prevaleção apenas com as chuvas, contudo o termo "rega" também surge como destaque. Contrariamente aos restantes clusters a batata apresenta maior importância que o milho.

Figura 14 - Nuvem de Palavras Recorrentes (Cluster 2)

O Cluster 3 revelou-se como o mais representado em termos de número de municípios, distribuindo-se tanto a Norte como a Sul, abrangendo zonas do Litoral e do Interior, o que indica uma menor influência de fatores físicos e ambientais na agricultura. A Figura 15 destaca as palavras mais predominantes neste cluster: milho, batata, vinho, rega, estrume e campo. Ao contrário do cluster 2, neste cluster o termo da "rega" surge com muito maior destaque que o de "sequeiro", sugerindo a presença de culturas com necessidades mais elevadas de água.

Figura 15 - Nuvem de Palavras Recorrentes (Cluster 3)

Por fim, o Cluster 4 apresenta uma distribuição irregular, destacando-se a ausência de municípios representativos na região nordeste da área de estudo, mas com uma concentração elevada na região Centro. As palavras que mais se destacaram neste cluster foram milho, vinha, cultura, sementeira, semente, feijão e lavradores, conforme ilustrado

na Figura 16. Importa referir que o termo "gado" aparece na nuvem de palavras com um tamanho relativamente visivel contrariamente ao observado nos restantes exemplos, isto pode significar que a criação de gado possui alguma importância para os municípios assinalados.

Figura 16 - Nuvem de Palavras Recorrentes (Cluster 4)

Posteriormente, foi realizada uma segunda análise de clusters considerando um conjunto de variáveis explicativas, i.e., os scores das culturas (Quadro 4) e diferentes categorias de população agrícola (total, com menos de 15 anos, ativa e com mais de 65 anos), com o objetivo de validar os resultados previamente obtidos. A Figura 17 ilustra a distribuição espacial dos clusters.

O Cluster 1 está concentrado principalmente na região central da área de estudo, particularmente nas Beiras (Litoral e Interior). De acordo com o modelo, as principais variáveis que definiram este cluster foram o milho, com cerca de 6%, seguido pela cevada e batata, ambas com 5,5%, e a população agrícola masculina em idade ativa, com 5,1%.

Por fim, o Cluster 4 tem um foco principal na região de Trás-os-Montes, mas também abrange alguns municípios dos distritos do Porto e Aveiro. As variáveis que melhor descrevem as características destas regiões foram o milho, com 7,1%, o feijão, com 6,7%, a batata, com 6%, e a vinha, com 4%. Em ambos os clusters predominaram as importâncias das culturas, assentes no milho, diferenciandas principalmente à subcultura explorada. O cluster 4 apresenta uma grande concentração na região do douro, evidenciando mais uma vez a influência que a vinha tem na região.

O Cluster 2 concentra-se mais na região norte da área de estudo, especialmente na região de Entre o Douro e Minho. As variáveis que melhor explicam a distribuição deste cluster são predominantemente demográficas, destacando-se o número de mulheres em idade ativa que trabalham na agricultura, com cerca de 10%, o total da população agrícola feminina, também com 10%, a população feminina na faixa etária de 12 a 14 anos que trabalha na agricultura, com 9,1%, e os homens na mesma faixa etária.

O Cluster 3 apresenta a menor expressão em termos de distribuição, sendo mais relevante na Beira Litoral. As variáveis que mais contribuíram para a definição deste cluster foram a população agrícola masculina total e a população masculina em idade ativa, ambas com 5,7%, a importância relativa da cultura da vinha, com 5,6%, e a população agrícola masculina com mais de 65 anos, com aproximadamente 5%.

Nestes dois clusters predominaram questões demográficas com espacial destaque para a divisão entre Homens mais velho e Mulheres mais jovens. A divisão destes dois clusters está intrinsecamente ligada à tradição agrícola do país uma vez que alguma cultura como a olivicultura e a vinicultura tendem a empregar mais homens, especialmente mais velhos, uma vez que estas atividades são desenvolvidas por famílias que evoluem de uma geração para a outra. Desta forma a experiência será valorizada e, aqui, os homens com idade mais avançada podem desempenhar tarefas de controlo e supervisão. Tal se verifica em alguns dos termos destacados no cluster 3. Relativamente ao cluster 2, as mulheres jovens e mulheres no geral desempenhavam funções onde fosse exigida uma maior destreza e agilidade, sendo isto especialmente importante na colheita de algumas culturas. É comum verificar mais mulheres nas culturas sazonais como a apanha dos frutos e nas de estufa (e.g., legumes) (Saltão e Moreira, 2006).

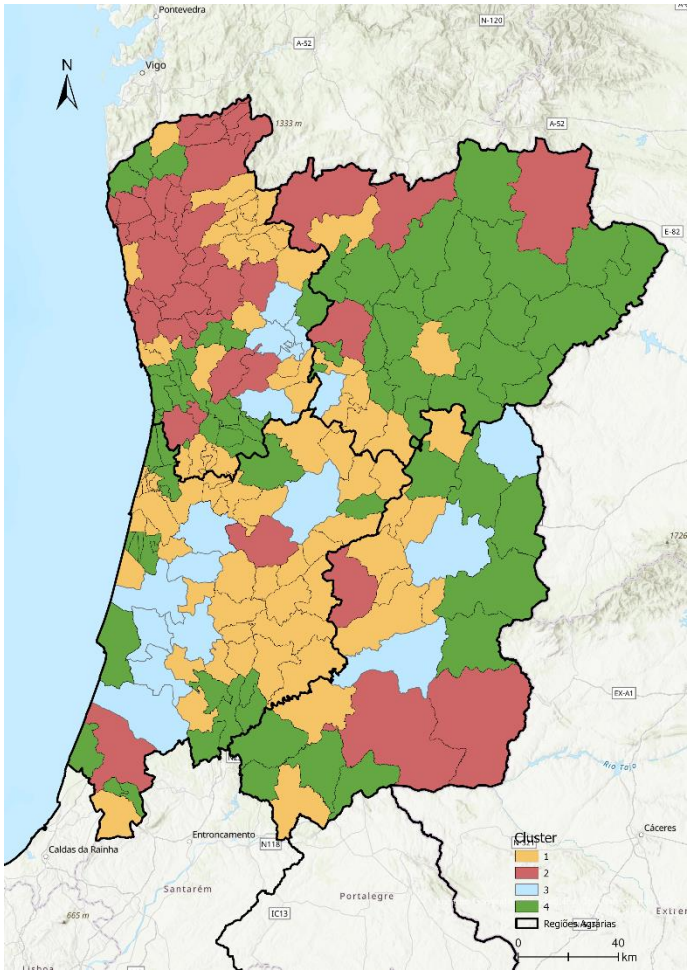


Figura 17 - Distribuição de Clusters com variáveis explicativas (K=4)

Em suma, o mapa da Figura 12 revela uma distribuição de clusters relativamente homogénea, com grupos próximos e contínuos, especialmente na região nordeste (Trás-os-Montes) e na zona central da área de estudo. Os clusters dos grupos 3 e 4 são predominantes, evidenciando uma significativa homogeneidade regional. Em contraste, as regiões pertencentes aos clusters 1 e 2 são menores e encontram-se concentradas em áreas mais específicas. Esta distribuição sugere que a análise, baseada unicamente no inquérito agrícola, favoreceu grandes agrupamentos regionais com características agrárias semelhantes, sendo o tipo de cultura o fator predominante na distribuição destes grupos. Outro ponto relevante para esta configuração é o facto de o inquérito ser rico em informação, incluindo descrições de variáveis físicas do território, o que também permite agrupar municípios com base em proximidades a rios, serras, entre outros elementos.

Por outro lado, o mapa da Figura 17 evidencia uma maior dispersão e fragmentação dos clusters. Aqueles que, no mapa anterior, apresentavam grandes blocos contínuos de municípios, agora estão mais fragmentados. Observa-se que, nos clusters 2 e 3, as características sociodemográficas assumem maior relevância do que o tipo de cultura. No entanto, é importante sublinhar que, apesar de no primeiro mapa os clusters parecerem focar-se principalmente no tipo de culturas, a análise considerou todo o conteúdo do inquérito, incluindo descrições quantitativas e qualitativas de diversos aspetos do setor agrícola, o que contribui para uma distribuição de clusters mais precisa.

Destaca-se ainda que a região de Trás-os-Montes é, de todas as regiões agrícolas analisadas, aquela com características agrárias mais homogéneas, sendo representada de forma bastante compacta em ambos os mapas. As principais características que definem os grupos nesta região indicam que o trigo e a vinha são elementos fundamentais na construção da sua identidade agrícola regional.

Capítulo 5 - Considerações Finais

Esta dissertação teve como principal objetivo explorar o uso da Inteligência Artificial (IA) em conjunto com os Sistemas de Informação Geográfica (SIG) para identificar e analisar as tendências agrícolas nos municípios das regiões agrárias da Beira Litoral, Beira Interior, Entre Douro e Minho, e Trás-os-Montes durante a década de 1950. A aplicação destas ferramentas permitiu extrair a informação desejada a partir de dados históricos, nomeadamente os Inquéritos Agrícolas e Florestais da década de 1950, possibilitando uma nova abordagem que oferece uma perspetiva renovada sobre as práticas agroecológicas da época.

Através desta dissertação, foi possível desenvolver uma análise exploratória inovadora que contribui para o campo das humanidades digitais, especialmente na combinação da inteligência artificial com os SIG. Ao utilizar estas ferramentas tecnológicas com dados históricos, foi possível superar algumas dificuldades inerentes à utilização de documentos antigos, facilitando a legibilidade e acessibilidade ao seu conteúdo. A principal contribuição para este campo foi a utilização conjunta do OCR e do ChatGPT, melhorando a extração de dados de fontes históricas e demonstrando como a IA pode aumentar a precisão e a qualidade dos dados, especialmente em grandes volumes.

Assim, no que respeita ao primeiro objetivo, foi criada uma metodologia sólida de recolha e tratamento de informação textual e cartográfica, que permitiu converter os dados históricos em formatos estruturados adequados para análise espacial. Os resultados indicam que o uso de IA, nomeadamente através de algoritmos de Natural Language Processing e Machine Learning, demonstrou um potencial significativo para automatizar e agilizar processos anteriormente manuais, comprovando a validade da Hipótese 1, que sugere que a IA melhora os métodos tradicionais de recolha e tratamento de dados históricos. Neste contexto, as ferramentas de OCR e a aplicação de técnicas de processamento de linguagem natural foram fundamentais para superar as dificuldades relacionadas com a qualidade dos inquéritos, que se encontravam deteriorados, comprometendo a legibilidade do texto. A utilização do ChatGPT, numa fase posterior à aplicação do OCR, revelou-se um ponto forte deste estudo, representando uma abordagem distinta em relação às práticas convencionais com ferramentas de OCR, abrindo caminho para avanços nesta área específica. O ChatGPT demonstrou ser uma

ferramenta excecionalmente eficaz, melhorando substancialmente a precisão dos dados textuais extraídos pelo OCR.

Relativamente ao segundo objetivo, a exploração dos algoritmos de IA mostrou-se particularmente útil na identificação e clusterização das especificidades agrícolas regionais. A análise permitiu verificar padrões de semelhança e distinção entre diferentes regiões do país, confirmando a Hipótese 2, de que os algoritmos permitem agrupar características agrícolas de forma eficaz. Concretamente, durante a análise do conteúdo dos inquéritos foi possível identificar que o milho era a principal cultura em 74% dos municípios, com a batata a ocupar a segunda posição. Outras culturas, como a cortiça, o centeio, o trigo e a vinha, embora menos predominantes, apresentavam uma expressão específica em certas regiões, no entanto, vale ressalvar que este trabalho enquadra-se numa investigação experimental e que a distribuição geográfica das principais culturas carece de posterior trabalho, como exemplo, o aparecimento da cortiça como cultura principal em Mogadouro e Alfândega da fé. A matriz de correlação revelou relações positivas entre culturas como o milho e o feijão, o milho e a batata, e a cevada e a aveia, sugerindo que estas culturas podem coexistir beneficamente. Por outro lado, surgiram correlações negativas entre o castanheiro e a amendoeira, que indicam possíveis desvantagens no seu cultivo conjunto. O índice de Shannon destacou a diversidade significativa de culturas, especialmente nos municípios da Beira Litoral e Beira Interior, enquanto em Trás-os-Montes foi observada uma menor diversidade, com uma produção agrícola mais focada em pouca diversidade agrícola. A análise de clusters mostrou que os municípios se agrupam de forma distinta, com os municípios de Trás-os-Montes a apresentar uma configuração agrária homogénea e bastante diferenciada das restantes regiões. Neste contexto, foi também possível observar que o milho, a batata, a vinha e o centeio foram as culturas de maior importância em várias regiões. Com destaque para o milho que se tornou amplamente predominante. Em Trás-os-Montes, o trigo e a vinha foram realçados como elementos essenciais da identidade agrícola regional. Além disso, a segunda análise de clusters revelou que a distribuição das culturas agrícolas não era apenas influenciada pelas condições edafoclimáticas, como o tipo de solo e a proximidade de recursos hídricos, mas também por características demográficas, como a composição da população agrícola. Por exemplo, na região do Douro, ambos os clusters destacaram a produção do vinho, sendo que nesta segunda análise confirmou-se que a tradição das culturas agrícolas e a população agrícola influenciavam o tipo de culturas plantadas. A

distribuição espacial dos clusters demonstrou que diferentes regiões possuíam características agrícolas específicas, refletidas nas suas culturas predominantes, evidenciando uma homogeneidade em algumas zonas e maior fragmentação noutras. Estes resultados permitem compreender como o contexto agroecológico do passado tem evoluído, oferecendo uma oportunidade para identificar possíveis mudanças futuras na agricultura.

Por fim, o terceiro objetivo foi atingido através da criação de um mapa agrícola da década de 1950, que sintetiza as informações cartográficas históricas recolhidas. Este mapa evidencia variações significativas na importância das culturas agrícolas entre as diferentes regiões, corroborando a Hipótese 3 sobre a diversidade regional nas práticas agrícolas. A análise espacial mostrou que a localização geográfica, as condições ambientais e a tradição cultural desempenham papéis decisivos na predominância de determinadas culturas em cada região.

Contudo, foram encontradas algumas limitações. O processo de digitalização dos inquéritos e a utilização do OCR foram processos muito demorados com as ferramentas de OCR atuais a apresentar uma precisão inferior a 75%, devido à deterioração dos documentos. O ChatGPT, apesar de melhorar significativamente os resultados, também implicou um processo de correção moroso, devido à limitação de caracteres e à inconsistência das respostas. Para trabalhos futuros, recomenda-se o desenvolvimento de um chatbot específico com instruções adequadas, capaz de lidar com maiores volumes de dados de uma só vez.

Concluindo, esta dissertação contribuiu para a aplicação de técnicas de IA no estudo de padrões históricos e agrícolas, destacando a relevância de métodos interdisciplinares para a análise de fenómenos complexos. A integração de IA e SIG proporcionou uma nova perspetiva sobre as dinâmicas agrícolas em Portugal durante a década de 1950, abrindo caminho para futuras investigações que explorem a interseção entre tecnologia e história. Os resultados desta dissertação apresentam uma nova forma de olhar para a gestão agrícola e territorial através de novos dados agrícolas úteis para enfrentar os desafios contemporâneos relacionados com a segurança alimentar, as alterações climáticas e a sustentabilidade. Uma sugestão para investigações futuras seria integrar esta análise com dados atuais, permitindo comparar a evolução das especificidades agroecológicas desde a década de 1950 até à atualidade e, assim, traçar

possíveis cenários futuros. Esta abordagem poderia revelar mudanças significativas nas dinâmicas de cultivo e uso do solo, influenciadas por fatores como alterações climáticas e mudanças nos padrões de consumo, oferecendo uma nova perspetiva para a formulação de políticas públicas agrícolas e ambientais. Por fim, sugere-se para futuros estudos a integração de diferentes algoritmos de clustering, como o clustering hierárquico e o DBSCAN, para revelar novos padrões ou identificar subgrupos e outliers.Calcular o EMQ da georreferenciação para verificar a viabilidade da mesma. A abordagem usada neste estudo pode ser expandida para outros temas com dados históricos, permitindo novas perspetivas sobre a evolução das práticas agrícolas e do uso do solo ao longo do tempo.

Bibliografia

Abro, a. A., talpur, m. S. H., & Jumani, a. K. (2023). Natural Language Processing Challenges and Issues: A Literature Review. Gazi University Journal of Science, 1-1.

Abubakar, H. D., Umar, M., & Bakale, M. A. (2022). Sentiment classification: Review of text vectorization methods: Bag of words, Tf-Idf, Word2vec and Doc2vec. SLU Journal of Science and Technology, 4(1 & 2), 27-33.

Alomari, A., Idris, N., Sabri, A. Q. M., & Alsmadi, I. (2022). Deep reinforcement and transfer learning for abstractive text summarization: A review. Computer Speech & Language, 71, 101276.

Ansari, F., Kohl, L., Giner, J., & Meier, H. (2021). Text mining for AI enhanced failure detection and availability optimization in production systems. *CIRP Annals*, *70*(1), 373-376.

Arndt, C., Farmer, W., Strzepek, K., & Thurlow, J. (2012). Climate change, agriculture and food security in Tanzania. Review of Development Economics, 16(3), 378-393.

Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 1027-1035).

Avillez, F. (2016). A Agricultura portuguesa. Fundação Francisco Manuel dos Santos.

Ayed, R., & Hanana, M. (2021). Artificial intelligence to improve the food and agriculture sector. Journal of Food Quality, 2021, 1-7.

Bação, Fernando & Painho, M.. (2003). Aspectos metodológicos da utilização de Data Mining no âmbito da Geografia. Finisterra: Revista Portuguesa de Geografia. 38. 10.18055/Finis1584.

Bates, M. (1995). Models of natural language understanding. Proceedings of the National Academy of Sciences, 92(22), 9977-9982.

Berchmans, D., & Kumar, S. S. (2014). Optical character recognition: an overview and an insight. In 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) (pp. 1361-1365). IEEE.

Berger, A., Caruana, R., Cohn, D., Freitag, D., & Mittal, V. (2000, July). Bridging the lexical chasm: statistical approaches to answer-finding. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 192-199).

Bhansali, A., Chandravadiya, A., Panchal, B. Y., Bohara, M. H., & Ganatra, A. (2022). Language Identification Using Combination of Machine Learning Algorithms and Vectorization Techniques. In 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 1329-1334). IEEE.

Biderman, M. T. C. (1996). Léxico e vocabulário fundamental. Alfa: revista de linguística, 40.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). New York: springer.

Bisol, L. (Ed.). (2005). Introdução a estudos de fonologia do português brasileiro. EdiPUCRS.

Bodenhamer, D. J., Corrigan, J., & Harris, T. M. (Eds.). (2010). *The spatial humanities: GIS and the future of humanities scholarship*. Indiana University Press.

Boivin, N., & Crowther, A. (2021). Mobilizing the past to shape a better Anthropocene. Nature Ecology & Evolution, 5(3), 273-284.

Braga, M., Araújo, S. D. S., Sales, H., Pontes, R., & Nunes, J. (2023). Portuguese Castanea sativa Genetic Resources: Characterization, Productive Challenges and Breeding Efforts. Agriculture, 13(8), 1629.

Brown, B., Chui, M., & Manyika, J. (2011). Are you ready for the era of 'big data'. McKinsey Quarterly, 4(1), 24-35.

Carvalho, F. P. (2006). Agriculture, pesticides, food security and food safety. Environmental science & policy, 9(7-8), 685-692.

Cawsey, A. J., Webber, B. L., & Jones, R. B. (1997). Natural language generation in health care. Journal of the American Medical Informatics Association, 4(6), 473-482.

Chang, I. C., Yu, T. K., Chang, Y. J., & Yu, T. Y. (2021). Applying text mining, clustering analysis, and latent dirichlet allocation techniques for topic classification of environmental education journals. Sustainability, 13(19), 10856.

Chen, D.Q., Preston, D.S., Swink, M., (2015). How the use of big data analytics affects value creation in supply chain management. J. Manage. Inform. Syst. 32 (4), 4–39. http://dx.doi.org/10.1080/07421222.2015.1138364.

Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). "Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review." *Computers and Electronics in Agriculture, 151*, 61-69.

Chou, C., & Lan, C. W. (2012). Changes in the annual range of precipitation under global warming. Journal of Climate, 25(1), 222-235.

Chou, C., Chiang, J. C., Lan, C. W., Chung, C. H., Liao, Y. C., & Lee, C. J. (2013). Increase in the range between wet and dry season precipitation. Nature Geoscience, 6(4), 263-267.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Copeland, B. Jack. "The turing test." Minds and Machines 10.4 (2000): 519-539.

Côrte-Real,N. (2022). BIG DATA & Analytics.Penguin Random House Grupo Editorial

Cui, M. (2020). Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance*, *1*(1), 5-8.

Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised learning. In Machine learning techniques for multimedia: case studies on organization and retrieval (pp. 21-49). Berlin, Heidelberg: Springer Berlin Heidelberg.

Danyal, M. M., Khan, S. S., Khan, M., Ullah, S., Ghaffar, M. B., & Khan, W. (2024). Sentiment analysis of movie reviews based on NB approaches using TF–IDF and count vectorizer. Social Network Analysis and Mining, 14(1), 1-15.

Delgado, J. A., Short Jr, N. M., Roberts, D. P., & Vandenberg, B. (2019). Big data analysis for sustainable agriculture on a geospatial cloud framework. Frontiers in Sustainable Food Systems, 3, 54.

Dobre, C., & Xhafa, F. (2014). Intelligent services for big data science. *Future generation computer systems*, *37*, 267-281.

Dodda, R., & Babu, A. S. (2024). Text document clustering using mayfly optimization algorithm with k-means technique. *Indonesian Journal of Electrical Engineering and Computer Science, 35*(2), 1099-1109. https://doi.org/10.11591/ijeecs.v35.i2.pp1099-1109

Doll, D. A., Andrade, J. F., & Serrano, P. (2021). Produção de amêndoa em Portugal: Tendências de plantação e desafios de produção num sector em desenvolvimento. AGRO.GES. https://youtu.be/ZqOmWf-bFRI

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big data–evolution, challenges and research agenda. International journal of information management, 48, 63-71.

Edwards, J. S., Duan, Y., & Robins, P. (2000). An analysis of expert systems for business decision making at different levels and in different roles. European Journal of Information Systems, 9(1), 36–46.

Eke, D. O. (2023). ChatGPT and the rise of generative AI: Threat to academic integrity?. Journal of Responsible Technology, 13, 100060.

Eliseu, A. (2014). Sintaxe do português. Leya.

Ertel, W. (2018). Introduction to artificial intelligence. Springer

FAO, 2017. The Future of Food and Agriculture - Trends and Challenges Rome.

Feldman, R., & Dagan, I. (1995, August). Knowledge Discovery in Textual Databases (KDT). In *KDD* (Vol. 95, pp. 112-117).

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., & Romero, C. (2019). Text mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(6), e1332.

Fichman, R. G., Dos Santos, B. L., & Zheng, Z. (2014). Digital innovation as a fundamental and powerful concept in the information systems curriculum. MIS quarterly, 38(2), 329-A15.

Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text mining methods and techniques. International Journal of Computer Applications, 85(17).

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International journal of information management, 35(2), 137-144.

Gatta, G., Arioti, E., & Bitelli, G. (2017). Geomatics science applied to cartographic heritage and archive sources: A new way to explore the XIXth century Gregorian Cadastre of Bologna (Italy), an ante-litteram 3D GIS. Journal of cultural heritage, 23, 68-76.

G.Almatar, M., Alazmi, H. S., Li, L., & Fox, E. A. (2020). Applying GIS and Text Mining Methods to Twitter Data to Explore the Spatiotemporal Patterns of Topics of Interest in Kuwait. ISPRS International Journal of Geo-Information, 9(12), 702.

Godfray, H.C.J., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence, D., Muir, J.F., Pretty, J., Robinson, S., Thomas, S.M., Toulmin, C., (2010). Food security: the challenge of feeding 9 billion people. Science (80-.) https://doi.org/10.1126/science.1185383.

Goyal, H., Sharma, C., & Joshi, N. (2017). An integrated approach of GIS and spatial data mining in big data. Int. J. Comput. Appl, 169(11), 1-6.

Gregory, I. N., & Ell, P. S. (2007). Historical GIS: technologies, methodologies, and scholarship (Vol. 39). Cambridge University Press.

Gregory, I. N., & Healey, R. G. (2007). Historical GIS: structuring, mapping and analysing geographies of the past. Progress in human geography, 31(5), 638-653.

Gregory, I. N., Bennett, C., Gilham, V. L., & Southall, H. R. (2002). The Great Britain Historical GIS Project: from maps to changing human geography. The Cartographic Journal, 39(1), 37-49.

Gregory, I. N., Kemp, K. K., & Mostern, R. (2001). Geographical Information and historical research: Current progress and future directions. History and Computing, 13(1), 7-23.

Gregory, I. N., & Geddes, A. (Eds.). (2014). Toward spatial humanities: Historical GIS and spatial history. Indiana University Press.

Halal, W. E. (2003) Artificial intelligence is almost here, On the Horizon - The Strategic Planning Resource for Education Professionals, Vol. 11, No. 2.

Han, J., Pei, J., & Kamber, M. (2011). Chapter 10 Cluster Analysis: Basic Concepts and Methods. Em Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Hegghammer, T. (2022). OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment. Journal of Computational Social Science, 5(1), 861-882.

Hey, T., Tansley S., & Tolle, K. (Eds.). (2009). The fourth paradigm: Data-intensive scientific discovery

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. Organizational Research Methods, 25(1), 114-146.

Holley, R. (2009). How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. D-Lib Magazine, 15(3/4).

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. Journal for Language Technology and Computational Linguistics, 20(1), 19-62.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. ACM computing surveys (CSUR), 31(3), 264-323.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651-666.

Jamiy, F. E., Daif, A., Azouazi, M., & Marzak, A. (2015). The potential and challenges of Big data-Recommendation systems next level application. *arXiv preprint arXiv:1501.03424*.

Jiang, Y., Li, X., Luo, H., Yin, S., & Kaynak, O. (2022). Quo vadis artificial intelligence. Discover Artificial Intelligence, 2(1), 4.

Kang, Y., Cai, Z., Tan, C. W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. Journal of Management Analytics, 7(2), 139-172.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., & Kumar, V. (2018). Machine learning for the geosciences: Challenges and opportunities. IEEE Transactions on Knowledge and Data Engineering, 31(8), 1544-1554

Keim, D. A., Panse, C., Sips, M., & North, S. C. (2004). Pixel based visual data mining of geo-spatial data. Computers & Graphics, 28(3), 327-344.

Knowles, A. K. (2005). Emerging trends in historical GIS. Historical Geography, 33(1), 7-13.

Knowles, A. K. (Ed.). (1999). Historical GIS: the spatial turn in social science history. Duke University Press.

Knowles, A. K., & Hillier, A. (2008). *Placing history: how maps, spatial data, and GIS are changing historical scholarship*. ESRI, Inc..′

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text mining in organizational research. *Organizational research methods*, *21*(3), 733-765.

Koistinen, M., Kettunen, K., & Kervinen, J. (2017). How to improve optical character recognition of historical Finnish newspapers using open source Tesseract OCR engine. Proc. of LTC, 279-283.

Krutilla, Z., & Kovari, A. (2022). The origin and primary areas of application of natural language processing. In 2022 IEEE 22nd International Symposium on Computational Intelligence and Informatics and 8th IEEE International Conference on Recent Achievements in Mechatronics, Automation, Computer Science and Robotics (CINTI-MACRo) (pp. 000293-000298). IEEE.

Kuriscak, E., Marsalek, P., Stroffek, J., & Toth, P. G. (2015). Biological context of Hebb learning in artificial neural networks, a review. Neurocomputing, 152, 27-35.

Lary, D. J. (2010). Artificial intelligence in geoscience and remote sensing. In Geoscience and Remote Sensing New Achievements. IntechOpen.

Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. Geoscience Frontiers, 7(1), 3-10.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196). PMLR.

Lee, R. (2011). The outlook for population growth. Science, 333(6042), 569-573.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674.

Liddy, E. D. (2001). Natural language processing.

Lu, Y. (2017). Industry 4.0: A survey on technologies, applications and open research issues. Journal of industrial information integration, 6, 1-10.

Mantas, J. (1986). An overview of character recognition methodologies. Pattern recognition, 19(6), 425-430.

Maynard, A. D. (2015). Navigating the fourth industrial revolution. Nature nanotechnology, 10(12), 1005-1006.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. AI magazine, 27(4), 12-12.

McDonald, D. D. (2010). Natural language generation. Handbook of Natural Language Processing, 2, 121–144.

Mennis, J., & Guo, D. (2009). Spatial data mining and geographic knowledge discovery—An introduction. Computers, Environment and Urban Systems, 33(6), 403-408.

Merghadi, A., Yunus, A. P., Dou, J., Whiteley, J., ThaiPham, B., Bui, D. T., ... & Abderrahmane, B. (2020). Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. Earth-Science Reviews, 207, 103225.

Misra, N. N., Dixit, Y., Al-Mallahi, A., Bhullar, M. S., Upadhyay, R., & Martynenko, A. (2020). IoT, big data, and artificial intelligence in agriculture and food industry. IEEE Internet of things Journal, 9(9), 6305-6324.

Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, *80*, 449-461.

Mithe, R., Indalkar, S., & Divekar, N. (2013). Optical character recognition. International journal of recent technology and engineering (IJRTE), 2(1), 72-75.

Mori, S., Suen, C. Y., & Yamamoto, K. (1992). Historical review of OCR research and development. Proceedings of the IEEE, 80(7), 1029-1058.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of big data*, *2*, 1-21.

Nagy, G. (1992). At the frontiers of OCR. Proceedings of the IEEE, 80(7), 1093-1100.

Nuckols, J. R., Ward, M. H., & Jarup, L. (2004). "Using geographic information systems for exposure assessment in environmental epidemiology studies." *Environmental Health Perspectives, 112*(9), 1007-1015.

Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, *219*(1), 103-119.

Portes, T. D. A. (1984). Aspectos ecofisiológicos do consórcio milho x feijão.

Rani, D., Kumar, R., & Chauhan, N. (2022). Study and comparision of vectorization techniques used in text classification. In 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

Renganathan, V. (2017). Text mining in biomedical domain with emphasis on document clustering. Healthcare informatics research, 23(3), 141-146.

Saltão,B., Moreira,J. (2006). Rotação de Culturas em Estufa. Direção Geral de Agricultura da Beira Interior

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513-523.

Sarker, I. H. (2021). Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. SN Computer Science, 2(6), 420.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN computer science, 2(3), 160.

Sarker, I. H. (2022). Ai-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems. SN Computer Science, 3(2), 158.

Shekhar, S., & Huang, Y. (2001). "Discovering spatial co-location patterns: A summary of results." *Proceedings of the 7th International Symposium on Advances in Spatial and Temporal Databases*.

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. IEEE access, 8, 80716-80727.

Tauger, M. B. (2010). *Agriculture in world history*. Routledge.

Thompson, P., McNaught, J., & Ananiadou, S. (2015). Customised OCR correction for historical medical text. In 2015 digital heritage (Vol. 1, pp. 35-42). IEEE.

Tomlinson, I.,(2013). Doubling food production to feed the 9 billion: a critical perspective on a key discourse of food security in the UK. J. Rural. Stud. 29, 81–90. https://doi. org/10.1016/j.jrurstud.2011.09.001.

Tóth, G., Hermann, T., Ravina da Silva, M., Montanarella, L., (2018). Monitoring soil for sustainable development and land degradation neutrality. Environ.Monit.Assess. 190, 57.

Trstenjak, B., Mikac, S., & Donko, D. (2014). KNN with TF-IDF based framework for text categorization. Procedia Engineering, 69, 1356-1364.

Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59(236), 433-460. https://doi.org/10.1093/mind/LIX.236.433

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. Information processing & management, 50(1), 104-112.

Viana, C. M., Freire, D., Abrantes, P., & Rocha, J. (2021). Evolution of agricultural production in Portugal during 1850–2018: A geographical and historical perspective. *Land*, *10*(8), 776.

Viana, C. M., Freire, D., Abrantes, P., Rocha, J., & Pereira, P. (2022). Agricultural land systems importance for supporting food security and sustainable development goals: A systematic review. *Science of the total environment*, *806*, 150718.

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. International Journal of Computer Science & Communication Networks, 5(1), 7-16.

Volk, M., Furrer, L., & Sennrich, R. (2011). Strategies for reducing and correcting OCR errors. In Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series (pp. 3-22). Springer Berlin Heidelberg.

Wall, K. (1986). A divisão sexual do trabalho na agricultura: Elementos para o seu estudo. Análise Social, 22(92/93), 661–668. https://www.jstor.org/stable/41010694

Wiedemann, G. (2013). Opening up to big data: Computer-assisted analysis of textual data in social sciences. *Historical Social Research/Historische Sozialforschung*, 332-357.

Witten, I. H. (2004). Text Mining.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., ... & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*, 1-37.

Yadav, S. S., Hegde, V. S., Habibi, A. B., Dia, M., & Verma, S. (2019). Climate change, agriculture and food security. Food security and climate change, 1-24.

Zhang, Weiwei & Zhai, Guangyu & Zhong, Binbin & Kong, Xiaoyi. (2024). Text Semantic Analysis Algorithm Based on LDA Model and Doc2vec. 10.3233/ATDE231269.

Zong, C., Xia, R., & Zhang, J. (2021). Text data mining (Vol. 711, p. 712). Singapore: Springer.

Anexos

Anexo 1 – stopwords adicionadas

'têm', 'áreas', 'ainda', 'importância', 'generalidade', 'pois', 'concelho', 'concelho,', 'algumas', 'pouco', 'forma', 'ano', 'área', 'concelho', 'bastante', 'devido', 'devidamente', 'a)', 'que,', 'maioria', 'média', 'alguns', 'menos', 'apenas', 'número', '1000', 'Quadro', 'fundas:', 'horas', 'consideráveis', 'concelho.', 'Vilarica,', '1000', 'Durante', 'terras', 'é,', 'quer', 'medianamente', 'cada', 'considerável', 'Vezes', 'vezes', 'quadro', 'Quadro', 'Fria,', 'b)', 'c)', 'Fria', '50', 'capacidade', 'B', 'Freixo', 'Espada', 'Miranda', 'Zona', 'Vilarica,', 'B', 'C', 'liga', 'caminho', 'Reparar', 'zona', 'freguesia', 'Castelo', 'Branco', 'Freguesias', 'sobretudo', 'anos', 'prédios', 'fim', 'pequenas', 'proprietário', 'grandes', 'duas', 'Vila', 'Serr', 'outro', 'km', 'km.', 'campo,', 'existem', 'Louriçal', 'Castelo', 'castelo', '2', 'Reparar', 'litros', 'vez', 'faz-se', 'condições', 'total', 'além', 'falta', 'outras', 'Vila', 'vila', 'ha', 'Reparar', 'freguesias', 'deve', 'outras', 'além', 'bem', 'Reparar ', 'proprietários', 'junto', 'juntar', '€.', 'seguintes', 'principais', 'vilarica,', 'proprietários', 'norte', 's.', 'norte,', 'Reparar', 'durante', 'pequenos', 'pequena', 'Freixo', 'deste', 'de', 'a', 'o', 'que', 'e', 'do', 'da', 'em', 'um', 'para', 'é', 'com', 'não', 'uma', 'os', 'no', 'se', 'na', 'por', 'mais', 'as', 'dos', 'como', 'mas', 'ao', 'ele', 'das', 'à', 'sua', 'ou', 'ser', 'quando', 'muito', 'há', 'nos', 'já', 'está', 'eu', 'também', 'só', 'pelo', 'pela', 'até', 'isso', 'ela', 'entre', 'era', 'depois', 'sem', 'mesmo', 'aos', 'ter', 'seus', 'quem', 'nas', 'me', 'esse', 'eles', 'estão', 'você', 'tinha', 'foram', 'essa', 'num', 'nem', 'suas', 'meu', 'às', 'minha', 'têm', 'numa', 'pelos', 'elas', 'havia', 'seja', 'qual', 'será', 'nós', 'tenho', 'lhe', 'deles', 'essas', 'esses', 'pelas', 'este', 'fosse', 'dele', 'tu', 'te', 'vocês', 'vos', 'lhes', 'meus', 'minhas', 'teu', 'tua', 'teus', 'tuas', 'nosso', 'nossa', 'nossos', 'nossas', 'dela', 'delas', 'esta', 'estes', 'estas', 'aquele', 'aquela', 'aqueles', 'aquelas', 'isto', 'aquilo', 'estou', 'está', 'estamos', 'estão', 'estive', 'esteve', 'estivemos', 'estiveram', 'estava', 'estávamos', 'estavam', 'estivera', 'estivéramos', 'esteja', 'estejamos', 'estejam',
'estivesse', 'estivéssemos', 'estivessem', 'estiver', 'estivermos', 'estiverem', 'hei', 'há', 'havemos', 'hão', 'houve', 'houvemos', 'houveram', 'houvera', 'houvéramos', 'haja', 'hajamos', 'hajam', 'houvesse',
'houvéssemos', 'houvessem', 'houver', 'houvermos', 'houverem', 'houverei', 'houverá', 'houveremos', 'houverão', 'houveria', 'houveríamos', 'houveriam', 'sou', 'somos', 'são', 'era', 'éramos', 'eram', 'fui',
'foi', 'fomos', 'foram', 'fora', 'fôramos', 'seja', 'sejamos', 'sejam', 'fosse', 'fôssemos', 'fossem', 'for', 'formos', 'forem', 'serei', 'será', 'seremos', 'serão', 'seria', 'seríamos', 'seriam', 'tenho', 'tem', 'temos', 'tém', 'tinha', 'tínhamos', 'tinham', 'tive', 'teve', 'tivemos', 'tiveram', 'tivera', 'tivéramos', 'tenha', 'tenhamos', 'tenham', 'tivesse', 'tivéssemos', 'tivessem', 'tiver', 'tivermos', 'tiverem','terei', 'terá', 'teremos', 'terão', 'teria', 'teríamos', 'teriam', 'região', 'reparar', 'se', 'todo', 'ano', 'normalmente', 'feita', 'toda', 're', 'Se', '-se', 'Cinta', 'caso', 'tipo', 'muitas', 'melhores', '10', '3', 'plena', 'locais', 'quantidades', 'Espada', 'valores', 'dois', 'assim', 'kg,', 'hectares.', 'sítio', 'vezes', 'preço', 'aumento', 'fazer', '15', 'principal', 'melhor', 'destacam-se', 'Quente,', 'B', 'anual', 'Vale', 'Fria,', 'C', 'maneira', 'extensão', 'vezes,', 'necessidades', 'utilizadas', '1º', 'qualquer', 'destinados', 'frequentemente', 'terra', '1.000', 'tanto', 'parece', 'quais', 'Dentre',
'Apesar', 'certa', 'expansão', 'aproximadamente', 'concelhos', 'seguintes:', '2º', 'ocorre', 'realizada', 'propriedade', 'poderiam', 'comum', 'preparação', 'enquanto', 'construção', 'interesse', 'vantajoso',
'certo', 'poderia', 'Produtos', 'médios', 'prática', 'fundas', 'vários', 'sido', 'após', 'superior', '1.500', 'Miranda', 'quanto', 'modo', 'podem', 'necessário', 'dentro', 'utilização', 'frequência,', 'uso',

'trabalhadores', 'hectare', 'Macedo', 'vale', 'interesse', 'faz', 'volta', 'porém,', 'todos', 'deveria',
'Zêzere,', 'todas', 'sistema', 'fazem', 'cursos', 'problema', 'considerar', 'porque', 'enquanto',
'deveriam', 'importante', 'antes', 'propriedade', 'vezes,', 'desta', 'serem', 'Covilhã',
'necessário', 'podem', 'Covilhã,', 'disso,', 'preparação', 'variedades', 'Conta', 'outros',
'distribuição',
'própria', 'Belmonte', 'conforme', 'três', 'quais', '5', '20', 'classe', 'relação', 'qualquer', 'devem',
'contra', '8', 'zonas', 'seguintes:', 'etc.', 'pois,', 'vão', 'aproveitamento',
'possibilidades', 'dada', 'material', 'propriedades', 'produtos', 'diferentes', 'necessário',
'prejuízos', 'zonas', '5', 'estrada', 'Vale', 'todos', 'construção', 'Branco', 'todas', 'outros', '20',
'devem',
'aproveitamento', 'desta', 'junto', 'existe', 'outra', 'podem', '4', 'possibilidades', 'muitos',
'valor', 'margens', '6', 'importante', 'base', 'aumentar', 'extensão', 'local', 'B', 'vezes,',
'interesse',
'linhas', 'pois,', 'maiores', 'Branco,', '8', 'Grémio', 'diretamente', 'sul', 'limite', '=', 'porém,',
'Feira', 'Ribeira', 'relação', 'C', 'restantes', 'regadio,', 'hectares', 'disso,', 'mal', '100',
'conforme', 'qualquer', 'isso,', 'seja,', 'junto', 'qualquer', 'valor', 'Rei', 'existe', 'linhas', 'fazem',
'concelhos', 'B', 'serem', 'outros', 'faz', 'C', '20', 'porque', 'caminhos', 'Zêzere', 'peso'
'aproveitamento', 'três', 'possibilidades', 'comum', 'acordo', 'vários', 'limite', 'produções',
'diferentes', 'solos', 'todos', 'divulgação', 'Volumes', 'kg:', 'escudos', 'sede', 'menor', '5',
'raramente',
'fontes', 'muitos', 'sido', 'Sertã,', 'Zêzere,', 'estrada', 'circunstância', 'através', 'único', 'Todos',
'melhorar', 'todas', 'outros', '1.000', 'interesse', 'sido', 'intervenção', 'B', 'inclui',
'necessidades', 'realizada', 'todas', 'anual', 'expansão', 'destinados', 'boa', 'poderia', 'venda',
'Quente.', 'Fria,', 'aproximadamente', 'assim,', 'enquanto', 'desses', 'todos', 'seguintes:',
'qualidade', 'significativa', 'variedades', 'podem', 'cuja', 'destaca-se', 'percentagem',
'Vilarica', 'qualquer', 'ocorre', 'Fria.', 'vantajoso', 'sede', 'meio', '100', 'C', 'benéfica', 'Torre',
'ponto', 'zonas', 'pertencentes', 'longo', 'localidades', 'tanto', 'Quente,', 'e)', 'cursos', 'escala',
'vários', 'características', 'preços', 'destacam-se', 'últimos', 'Freixo', '5', 'Espada'
'interesse', 'conforme', 'realizada', 'estrada', 'incluem', 'variedades', 'zonas', 'disso,', 'menor',
'base', 'anos,', 'Nacional', 'estação', 'local', 'enquanto', 'todos', 'contra', 'destinadas',
'outros', 'atualmente', 'Produtos', 'B', 'após', 'destinada', 'C', 'aproveitamento', 'existentes',
'Quanto', 'extensão', '(S.', 'sob', 'desses', 'hectares', 'ligação', 'valor', 'Cinta', 'todas',
'sobre', 'metade', 'quais', 'meio', 'variedade', 'acordo', 'desenvolvimento', 'Zêzere', 'Branco',
'Covilhã', 'feito','atualmente'
,'exemplo,','transporte','B','rio','C','delgadas:','povoações','Dentre','litros.','baixa',',','lugar','primeir
a','seguir','primeiro','vai','portanto,','Beira','tipos','regime','30','segunda','tendo','feitas','aqui','dad
os','pouca','consumo','trabalho','boas','desde','25','interessante','segundo','40','hábito','região.','l
imites','200','poucos','contrário,','toneladas','destas','dado','certas','B','matéria','poucas','dessa',
'dá''B','C','Apesar','frequência','importância,','Dentre','rios','I','conta','doenças','final','trabalho','a
lgum','600','apesar','desde','pragas','exclusivamente','convenientemente','conveniente','quais
quer','tal','Quente.','II','realizadas','reduzida','delgadas','cuidados','trabalhos','muita','litros,','D',
Cultura','possível','Bragança','dias','Macedo','altitude','principalmente,','boas','Miranda','arado'
,'Sabor,','também,','vantagens','populações','matéria','Quente','várias','importante','Cantanhe
de','estreme','mencionado','extensa','Grémio','diversos','Verificase','limpeza','considerada','ne
gociante','obra','leste','abundante','exemplo','Lagoa','locais','se','cultura','terreno','Cultura','Terr
eno','terrenos','culturas','&','produção','águas','-se','se','Se','água'

Anexo 2 – scores de culturas por munícipio

		Documentos						
		Alvaiazere	Amarante	Amares	Anadia	Ansiao	ArcosdeValdevez	Arganil
Culturas	**milho**	8.15	142.58	59.07	45.83	61.11	47.87	48.89
	batata	4.02	50.31	24.15	45.28	36.22	10.06	32.20
	vinha	4.07	17.31	17.31	43.79	20.37	7.13	21.39
	azeite	4.33	15.16	7.58	14.07	10.83	6.50	8.66
	arroz	1.19	1.19	2.39	26.28	2.39	2.39	2.39
	centeio	0.00	69.26	23.42	3.06	10.18	15.28	27.50
	trigo	2.00	26.00	19.00	35.00	29.00	13.00	18.00
	nabo	0.00	18.81	0.00	3.13	1.57	6.27	0.00
	feijão	2.05	57.38	34.84	9.22	28.69	15.37	22.54
	grão	0.00	10.55	9.38	7.04	17.59	1.17	21.11
	aveia	1.23	6.16	0.00	20.95	3.70	1.23	7.39
	oliveira	1.06	12.75	7.44	17.01	26.57	3.19	17.01
	amêndoa	0.00	3.10	0.00	0.00	0.00	0.00	0.00
	cevada	1.26	1.26	0.00	5.02	1.26	0.00	8.79
	chícharo	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cortiça	0.00	0.00	0.00	0.00	0.00	0.00	14.28
	noz	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	castanha	0.00	3.85	0.00	0.00	0.00	11.55	5.78
	avelã	0.00	0.00	0.00	0.00	0.00	3.80	0.00
	tremoço	2.95	10.32	1.47	0.00	7.37	0.00	2.95
	laranja	0.00	6.50	26.02	0.00	0.00	8.67	2.17
	limão	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mel	4.96	7.44	2.48	2.48	6.20	3.72	3.72
	aguardente	0.00	4.23	1.41	8.45	0.00	0.00	2.82
	pêssego	0.00	6.11	0.00	0.00	0.00	3.06	0.00
	pêra	0.00	0.00	0.00	0.00	0.00	0.00	6.42
	maçã	0.00	4.68	2.34	0.00	0.00	0.00	0.00

		Documentos						
		Armamar	Arouca	Aveiro	Baião	Barcelos	Batalha	Belmonte
Culturas	**milho**	42.78	24.44	72.31	105.92	78.42	31.57	45.83
	batata	26.16	7.04	53.32	41.25	39.24	18.11	33.20
	vinha	42.78	10.18	26.48	15.28	12.22	13.24	15.28
	azeite	6.50	2.17	1.08	15.16	6.50	8.66	10.83
	arroz	4.78	0.00	25.09	1.19	3.58	2.39	2.39
	centeio	23.42	2.04	8.15	56.02	21.39	3.06	37.68
	trigo	27.00	1.00	20.00	30.00	17.00	23.00	15.00
	nabo	9.40	0.00	26.64	9.40	1.57	0.00	4.70
	feijão	13.32	6.15	39.96	31.77	50.21	10.25	18.44
	grão	5.86	0.00	1.17	7.04	11.73	0.00	4.69
	aveia	7.39	0.00	12.32	1.23	27.11	6.16	2.46
	oliveira	17.01	3.19	9.57	17.01	7.44	9.57	11.69
	amêndoa	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cevada	18.84	0.00	12.56	15.07	0.00	6.28	2.51
	chícharo	3.90	0.00	0.00	0.00	0.00	0.00	0.00
	cortiça	2.04	0.00	0.00	2.04	0.00	0.00	2.04
	noz	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	castanha	1.93	1.93	0.00	0.00	1.93	0.00	1.93
	avelã	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	tremoço	19.17	0.00	10.32	4.42	0.00	2.95	1.47
	laranja	0.00	0.00	0.00	6.50	0.00	0.00	0.00
	limão	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mel	2.48	0.00	1.24	2.48	3.72	2.48	3.72
	aguardente	1.41	0.00	1.41	5.63	1.41	0.00	1.41
	pêssego	0.00	0.00	0.00	3.06	0.00	0.00	0.00
	pêra	6.42	0.00	0.00	0.00	0.00	0.00	0.00
	maçã	0.00	0.00	0.00	0.00	0.00	0.00	0.00

		Documentos						
		Boticas	Braga	Braganca	Cabeceiras de basto	Caminha	Cantanhede	Carrazeda
Culturas	**milho**	43.79	7.13	5.09	30.55	24.44	71.29	13.24
	batata	65.40	5.03	30.18	22.13	19.12	37.23	47.29
	vinha	25.46	7.13	6.11	11.20	10.18	41.76	27.50
	azeite	11.91	3.25	3.25	7.58	4.33	22.73	11.91
	arroz	3.58	0.00	1.19	1.19	1.19	37.03	8.36
	centeio	49.90	6.11	20.37	20.37	17.31	8.15	31.57
	trigo	3.00	2.00	23.00	13.00	8.00	28.00	22.00
	nabo	0.00	0.00	0.00	0.00	0.00	0.00	1.57
	feijão	13.32	0.00	1.02	24.59	8.20	35.86	10.25
	grão	3.52	0.00	2.35	5.86	2.35	12.90	7.04
	aveia	0.00	0.00	0.00	2.46	1.23	16.02	6.16
	oliveira	1.06	3.19	3.19	12.75	5.31	13.82	12.75
	amêndoa	0.00	0.00	0.00	0.00	0.00	0.00	9.31
	cevada	3.77	0.00	1.26	7.53	1.26	15.07	6.28
	chícharo	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cortiça	14.28	10.20	0.00	0.00	0.00	2.04	46.91
	noz	3.53	0.00	0.00	0.00	0.00	0.00	7.07
	castanha	9.63	0.00	9.63	0.00	0.00	0.00	21.18
	avelã	3.80	0.00	0.00	0.00	0.00	0.00	0.00
	tremoço	0.00	1.47	2.95	1.47	0.00	2.95	0.00
	laranja	0.00	0.00	0.00	0.00	0.00	4.34	8.67
	limão	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mel	3.72	1.24	0.00	3.72	6.20	3.72	4.96
	aguardente	7.04	0.00	0.00	0.00	1.41	2.82	9.86
	pêssego	0.00	0.00	0.00	0.00	0.00	3.06	0.00
	pêra	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	maçã	2.34	0.00	0.00	0.00	2.34	2.34	7.02

		Documentos						
		CarregalDoSal	CastanheiraDePera	CasteloBranco	CasteloDePaiva	CastroDaire	CeloricaDaBeira	CeloricaDeBasto
Culturas	**milho**	50.92	49.90	21.39	19.35	59.07	49.90	113.05
	batata	17.10	21.13	4.02	8.05	27.17	58.35	44.27
	vinha	14.26	10.18	2.04	12.22	17.31	15.28	13.24
	azeite	6.50	9.74	8.66	4.33	10.83	14.07	35.73
	arroz	1.19	2.39	1.19	3.58	2.39	2.39	1.19
	centeio	11.20	12.22	19.35	1.02	35.65	52.96	50.92
	trigo	10.00	9.00	13.00	2.00	17.00	10.00	38.00
	nabo	3.13	1.57	0.00	0.00	4.70	1.57	0.00
	feijão	20.49	24.59	11.27	2.05	22.54	14.35	33.81
	grão	28.14	11.73	0.00	1.17	21.11	14.07	17.59
	aveia	3.70	0.00	2.46	1.23	3.70	1.23	2.46
	oliveira	24.45	9.57	9.57	7.44	14.88	9.57	23.38
	amêndoa	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cevada	7.53	6.28	1.26	0.00	3.77	1.26	12.56
	chícharo	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cortiça	0.00	2.04	0.00	2.04	0.00	4.08	0.00
	noz	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	castanha	1.93	15.40	0.00	0.00	0.00	1.93	0.00
	avelã	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	tremoço	5.90	4.42	4.42	0.00	1.47	0.00	2.95
	laranja	0.00	0.00	2.17	4.34	2.17	0.00	6.50
	limão	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mel	7.44	3.72	4.96	4.96	4.96	0.00	2.48
	aguardente	1.41	0.00	1.41	1.41	0.00	0.00	4.23
	pêssego	0.00	0.00	3.06	6.11	0.00	0.00	0.00
	pêra	3.21	60.98	0.00	3.21	0.00	0.00	0.00
	maçã	0.00	0.00	2.34	0.00	0.00	0.00	7.02

		Documentos						
		Chaves	Cinfaes	Coimbra	CondeixaANova	Covilha	Espinho	Esposende
↻	**milho**	12.22	87.59	103.88	66.20	57.03	11.20	63.14

	batata	27.17	45.28	54.33	36.22	40.24	4.02	50.31
	vinha	14.26	39.72	43.79	28.52	25.46	2.04	12.22
	azeite	3.25	16.24	10.83	14.07	2.17	0.00	12.99
	arroz	0.00	1.19	74.06	35.84	2.39	0.00	2.39
	centeio	16.30	61.11	10.18	9.17	52.96	0.00	20.37
	trigo	17.00	15.00	39.00	40.00	21.00	4.00	16.00
	nabo	0.00	0.00	3.13	0.00	4.70	0.00	20.37
	feijão	3.07	30.74	39.96	27.67	18.44	3.07	46.11
	grão	0.00	7.04	63.32	34.01	7.04	0.00	4.69
	aveia	2.46	1.23	1.23	17.25	3.70	1.23	17.25
	oliveira	2.13	43.58	45.71	25.51	14.88	1.06	6.38
	amêndoa	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cevada	1.26	2.51	2.51	8.79	6.28	0.00	8.79
	chícharo	0.00	0.00	0.00	3.90	0.00	0.00	0.00
	cortiça	0.00	2.04	0.00	0.00	2.04	4.08	0.00
	noz	0.00	3.53	0.00	0.00	0.00	0.00	0.00
	castanha	0.00	25.03	0.00	0.00	1.93	0.00	0.00
	avelã	0.00	3.80	0.00	0.00	0.00	0.00	0.00
	tremoço	2.95	2.95	0.00	5.90	7.37	0.00	0.00
	laranja	0.00	2.17	4.34	0.00	0.00	0.00	0.00
	limão	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mel	1.24	2.48	7.44	7.44	3.72	2.48	3.72
	aguardente	8.45	2.82	0.00	0.00	0.00	0.00	2.82
	pêssego	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	pêra	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	maçã	0.00	2.34	0.00	0.00	0.00	0.00	0.00

Documentos

		Estarreja	Fafe	Felgueiras	Figueira De Castelo Rodrigo	FigueiraDaFoz	FigueiroDosVinhos	FornosDeAlgodres
Culturas	**milho**	61.11	50.92	62.13	6.11	70.27	36.66	53.98
	batata	19.12	28.17	24.15	10.06	41.25	12.07	32.20
	vinha	9.17	25.46	15.28	9.17	32.59	6.11	11.20
	azeite	1.08	9.74	3.25	5.41	10.83	0.00	17.32
	arroz	10.75	3.58	1.19	1.19	96.76	0.00	2.39
	centeio	8.15	34.63	23.42	14.26	10.18	8.15	29.54
	trigo	12.00	15.00	15.00	20.00	30.00	5.00	16.00
	nabo	0.00	0.00	3.13	0.00	1.57	0.00	3.13
	feijão	19.47	26.64	27.67	2.05	49.19	15.37	24.59
	grão	0.00	11.73	3.52	1.17	17.59	37.52	11.73
	aveia	14.79	0.00	0.00	1.23	19.72	0.00	2.46
	oliveira	3.19	9.57	0.00	6.38	10.63	7.44	9.57
	amêndoa	0.00	0.00	0.00	6.21	0.00	0.00	0.00
	cevada	7.53	0.00	0.00	3.77	17.58	5.02	6.28
	chícharo	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cortiça	0.00	0.00	0.00	2.04	0.00	0.00	0.00
	noz	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	castanha	0.00	1.93	0.00	3.85	0.00	1.93	5.78
	avelã	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	tremoço	1.47	1.47	8.85	1.47	0.00	1.47	1.47
	laranja	0.00	4.34	2.17	4.34	0.00	0.00	0.00
	limão	0.00	0.00	4.15	0.00	0.00	0.00	0.00
	mel	3.72	1.24	3.72	3.72	0.00	0.00	1.24
	aguardente	2.82	0.00	2.82	8.45	2.82	0.00	0.00

	pêssego	0.00	0.00	6.11	0.00	0.00	0.00	0.00
	pêra	0.00	0.00	9.63	0.00	0.00	0.00	0.00
	maçã	0.00	0.00	7.02	0.00	0.00	0.00	0.00

		Documentos						
		FreixodeEspadaaCinta	Fundão	Gois	Gondomar	Gouveia	Guarda	Guimarães
Culturas	**milho**	9.17	78.42	57.03	17.31	58.05	86.57	84.53
	batata	13.08	41.25	24.15	10.06	28.17	31.19	35.21
	vinha	12.22	27.50	24.44	8.15	28.52	33.61	23.42
	azeite	16.24	24.90	19.49	1.08	9.74	11.91	8.66
	arroz	1.19	3.58	0.00	0.00	2.39	2.39	1.19
	centeio	24.44	27.50	18.33	3.06	51.94	76.38	38.70
	trigo	32.00	48.00	13.00	1.00	15.00	16.00	18.00
	nabo	0.00	0.00	1.57	3.13	1.57	3.13	6.27
	feijão	4.10	43.04	27.67	6.15	33.81	45.09	46.11
	grão	5.86	17.59	3.52	1.17	12.90	14.07	24.62
	aveia	7.39	1.23	7.39	1.23	3.70	6.16	0.00
	oliveira	10.63	13.82	14.88	5.31	15.94	20.20	8.50
	amêndoa	3.10	0.00	0.00	0.00	0.00	0.00	0.00
	cevada	7.53	7.53	15.07	0.00	7.53	8.79	0.00
	chícharo	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cortiça	38.75	40.79	0.00	0.00	2.04	2.04	0.00
	noz	0.00	0.00	0.00	0.00	3.53	0.00	0.00
	castanha	9.63	23.11	0.00	1.93	1.93	26.96	0.00
	avelã	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	tremoço	0.00	11.80	5.90	0.00	5.90	5.90	5.90
	laranja	13.01	0.00	0.00	0.00	0.00	0.00	2.17
	limão	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mel	0.00	7.44	3.72	2.48	1.24	4.96	1.24
	aguardente	5.63	11.27	11.27	0.00	0.00	1.41	0.00
	pêssego	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	pêra	0.00	0.00	0.00	0.00	0.00	6.42	0.00

	maçã	0.00	2.34	0.00	0.00	0.00	0.00	0.00

		Documentos						
		IdanhaANova	Ilhavo	Lamego	Leiria	Lousa	Lousada	Mação
Culturas	**milho**	22.41	48.89	55.00	49.90	62.13	17.31	64.16
	batata	7.04	43.26	40.24	5.03	26.16	8.05	17.10
	vinha	3.06	9.17	57.03	30.55	14.26	12.22	18.33
	azeite	11.91	2.17	5.41	3.25	19.49	2.17	8.66
	arroz	4.78	3.58	3.58	9.56	0.00	3.58	51.37
	centeio	20.37	11.20	35.65	2.04	10.18	9.17	29.54
	trigo	28.00	18.00	27.00	12.00	24.00	7.00	43.00
	nabo	1.57	3.13	1.57	0.00	0.00	0.00	4.70
	feijão	9.22	29.72	21.52	17.42	36.89	10.25	33.81
	grão	4.69	7.04	9.38	12.90	3.52	0.00	15.24
	aveia	7.39	20.95	4.93	6.16	7.39	0.00	24.65
	oliveira	14.88	3.19	20.20	15.94	7.44	2.13	11.69
	amêndoa	0.00	0.00	3.10	3.10	0.00	0.00	0.00
	cevada	3.77	20.09	6.28	0.00	15.07	0.00	21.35
	chícharo	0.00	0.00	0.00	7.81	0.00	0.00	0.00
	cortiça	2.04	0.00	0.00	4.08	0.00	12.24	2.04
	noz	0.00	0.00	0.00	0.00	0.00	3.53	0.00
	castanha	0.00	0.00	7.70	3.85	0.00	3.85	0.00
	avelã	0.00	0.00	3.80	0.00	0.00	0.00	0.00
	tremoço	2.95	2.95	1.47	4.42	5.90	0.00	4.42
	laranja	2.17	0.00	0.00	0.00	0.00	0.00	0.00
	limão	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mel	7.44	0.00	0.00	0.00	3.72	2.48	7.44
	aguardente	1.41	0.00	4.23	4.23	9.86	4.23	1.41
	pêssego	0.00	0.00	3.06	0.00	3.06	0.00	0.00
	pêra	0.00	0.00	3.21	0.00	0.00	0.00	0.00
	maçã	0.00	0.00	2.34	0.00	2.34	0.00	0.00

		Documentos						
		Macedo de Cavaleiros	Maia	Mangualde	Manteigas	Marco de Canavezes	MarinhaGrande	Matosinhos
Culturas	**milho**	3.06	60.09	56.02	57.03	67.22	18.33	91.66
	batata	21.13	28.17	47.29	24.15	33.20	5.03	37.23
	vinha	5.09	9.17	16.30	25.46	12.22	8.15	10.18
	azeite	6.50	2.17	4.33	7.58	12.99	2.17	3.25
	arroz	2.39	2.39	1.19	1.19	1.19	4.78	2.39
	centeio	29.54	20.37	33.61	53.98	33.61	2.04	27.50
	trigo	23.00	19.00	3.00	4.00	11.00	2.00	19.00
	nabo	3.13	10.97	10.97	3.13	1.57	0.00	0.00
	feijão	1.02	16.40	15.37	33.81	36.89	7.17	22.54
	grão	1.17	19.93	7.04	9.38	5.86	0.00	37.52
	aveia	0.00	14.79	0.00	2.46	7.39	3.70	17.25
	oliveira	9.57	2.13	6.38	6.38	11.69	3.19	0.00
	amêndoa	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cevada	3.77	7.53	32.65	5.02	10.05	3.77	16.32
	chícharo	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cortiça	16.32	0.00	0.00	2.04	0.00	0.00	0.00
	noz	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	castanha	11.55	0.00	11.55	11.55	3.85	0.00	0.00
	avelã	0.00	0.00	11.39	0.00	0.00	0.00	0.00
	tremoço	1.47	2.95	0.00	0.00	5.90	1.47	0.00
	laranja	0.00	0.00	0.00	0.00	8.67	0.00	2.17
	limão	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mel	2.48	2.48	0.00	2.48	3.72	0.00	0.00
	aguardente	1.41	0.00	2.82	2.82	2.82	4.23	2.82
	pêssego	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	pêra	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	maçã	0.00	0.00	4.68	0.00	0.00	0.00	0.00

		Documentos						
		Mealhada	Meda	Melgaço	MesaoFrio	Mira	MirandaDoCorvo	MirandaDoDouro
Culturas	**milho**	72.31	36.66	39.72	27.50	57.03	67.22	1.02
	batata	54.33	31.19	31.19	5.03	40.24	35.21	19.12
	vinha	52.96	15.28	15.28	18.33	17.31	24.44	13.24
	azeite	21.65	10.83	1.08	9.74	4.33	8.66	7.58
	arroz	4.78	1.19	1.19	0.00	20.31	0.00	1.19
	centeio	9.17	29.54	25.46	3.06	10.18	10.18	28.52
	trigo	24.00	26.00	10.00	8.00	26.00	20.00	31.00
	nabo	1.57	10.97	0.00	0.00	3.13	4.70	1.57
	feijão	17.42	21.52	12.30	4.10	32.79	35.86	2.05
	grão	9.38	7.04	9.38	3.52	3.52	8.21	1.17
	aveia	25.88	1.23	0.00	1.23	13.56	3.70	0.00
	oliveira	19.13	15.94	0.00	20.20	3.19	13.82	11.69
	amêndoa	0.00	6.21	0.00	0.00	0.00	0.00	12.42
	cevada	21.35	17.58	7.53	0.00	15.07	12.56	1.26
	chícharo	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cortiça	2.04	2.04	0.00	12.24	2.04	0.00	36.71
	noz	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	castanha	1.93	1.93	0.00	0.00	0.00	0.00	9.63
	avelã	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	tremoço	0.00	5.90	8.85	0.00	1.47	8.85	0.00
	laranja	4.34	0.00	0.00	0.00	0.00	2.17	0.00
	limão	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mel	1.24	2.48	2.48	6.20	0.00	3.72	0.00
	aguardente	9.86	0.00	0.00	5.63	0.00	5.63	1.41
	pêssego	0.00	0.00	0.00	0.00	0.00	3.06	0.00
	pêra	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	maçã	0.00	0.00	0.00	2.34	0.00	2.34	0.00

		Documentos						
		Mirandela	Mogadouro	Moimenta da Beira	Monção	Mondim de Basto	Montalegre	Montemor-o-Velho
Culturas	**milho**	5.09	0.00	39.72	43.79	2.04	20.37	69.26
	batata	18.11	17.10	24.15	28.17	0.00	57.35	31.19
	vinha	8.15	10.18	16.30	11.20	0.00	5.09	19.35
	azeite	17.32	7.58	14.07	11.91	0.00	7.58	14.07
	arroz	1.19	1.19	2.39	0.00	0.00	0.00	76.45
	centeio	19.35	25.46	26.48	28.52	1.02	25.46	3.06
	trigo	32.00	27.00	14.00	11.00	1.00	2.00	25.00
	nabo	0.00	0.00	0.00	0.00	0.00	0.00	12.54
	feijão	3.07	1.02	17.42	24.59	0.00	2.05	43.04
	grão	3.52	2.35	4.69	3.52	0.00	4.69	7.04
	aveia	2.46	0.00	0.00	0.00	0.00	7.39	17.25
	oliveira	3.19	9.57	11.69	2.13	0.00	2.13	13.82
	amêndoa	0.00	15.52	0.00	0.00	0.00	3.10	0.00
	cevada	2.51	2.51	3.77	3.77	0.00	0.00	11.30
	chícharo	3.90	0.00	0.00	0.00	0.00	0.00	11.71
	cortiça	26.51	36.71	0.00	0.00	0.00	10.20	0.00
	noz	0.00	0.00	0.00	0.00	0.00	3.53	0.00
	castanha	1.93	19.25	9.63	0.00	0.00	9.63	0.00
	avelã	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	tremoço	2.95	0.00	1.47	0.00	0.00	0.00	1.47
	laranja	0.00	0.00	0.00	0.00	0.00	0.00	6.50
	limão	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mel	0.00	1.24	3.72	4.96	1.24	1.24	0.00
	aguardente	1.41	1.41	0.00	5.63	1.41	5.63	4.23
	pêssego	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	pêra	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	maçã	0.00	0.00	0.00	0.00	0.00	0.00	0.00

		Documentos						
		Mortagua	Murca	Murtosa	Nelas	Oleiros	OliveiradeAzemeis	OliveiraDeFrades
Culturas	**milho**	52.96	8.15	57.03	32.59	33.61	74.35	62.13
	batata	21.13	9.06	18.11	13.08	24.15	28.17	7.04
	vinha	9.17	17.31	4.07	21.39	17.31	15.28	12.22
	azeite	12.99	1.08	1.08	7.58	9.74	5.41	9.74
	arroz	3.58	0.00	0.00	2.39	1.19	10.75	0.00
	centeio	14.26	8.15	13.24	14.26	20.37	9.17	24.44
	trigo	28.00	6.00	10.00	12.00	15.00	10.00	1.00
	nabo	0.00	0.00	0.00	1.57	3.13	0.00	1.57
	feijão	21.52	7.17	22.54	16.40	17.42	29.72	28.69
	grão	1.17	1.17	0.00	23.45	1.17	4.69	4.69
	aveia	13.56	2.46	17.25	2.46	3.70	2.46	4.93
	oliveira	4.25	5.31	0.00	11.69	20.20	30.82	37.20
	amêndoa	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cevada	5.02	0.00	25.11	7.53	7.53	7.53	1.26
	chícharo	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cortiça	0.00	0.00	0.00	0.00	0.00	2.04	0.00
	noz	0.00	0.00	0.00	3.53	0.00	0.00	0.00
	castanha	0.00	7.70	0.00	0.00	0.00	0.00	1.93
	avelã	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	tremoço	2.95	1.47	1.47	0.00	1.47	2.95	0.00
	laranja	0.00	0.00	0.00	0.00	0.00	0.00	4.34
	limão	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mel	0.00	1.24	1.24	2.48	1.24	2.48	1.24
	aguardente	1.41	0.00	1.41	2.82	2.82	1.41	2.82
	pêssego	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	pêra	0.00	0.00	0.00	3.21	0.00	0.00	0.00
	maçã	0.00	0.00	2.34	0.00	0.00	0.00	2.34

		Documentos						
		OliveiraDoBairro	OliveiraDoHospital	Ovar	Paços de Ferreira	Pampilhosa da Serra	Paredes de Coura	Paredes
Culturas	**milho**	42.78	55.00	8.15	22.41	61.11	15.28	53.98
	batata	32.20	33.20	4.02	8.05	30.18	19.12	34.21
	vinha	17.31	2.04	0.00	18.33	20.37	13.24	7.13
	azeite	5.41	10.83	0.00	0.00	17.32	0.00	4.33
	arroz	35.84	3.58	9.56	0.00	2.39	0.00	2.39
	centeio	3.06	11.20	2.04	18.33	24.44	6.11	33.61
	trigo	7.00	12.00	3.00	7.00	11.00	10.00	28.00
	nabo	0.00	0.00	1.57	0.00	4.70	1.57	1.57
	feijão	19.47	17.42	3.07	12.30	31.77	10.25	32.79
	grão	1.17	5.86	2.35	2.35	4.69	3.52	2.35
	aveia	7.39	4.93	1.23	0.00	1.23	0.00	6.16
	oliveira	42.52	24.45	1.06	0.00	8.50	1.06	3.19
	amêndoa	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cevada	6.28	5.02	0.00	0.00	10.05	0.00	0.00
	chícharo	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cortiça	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	noz	0.00	14.14	0.00	0.00	0.00	0.00	0.00
	castanha	0.00	0.00	0.00	0.00	1.93	0.00	0.00
	avelã	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	tremoço	0.00	16.22	0.00	0.00	2.95	0.00	5.90
	laranja	2.17	0.00	0.00	0.00	0.00	0.00	2.17
	limão	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mel	2.48	0.00	0.00	0.00	3.72	0.00	3.72
	aguardente	7.04	1.41	0.00	0.00	2.82	0.00	1.41

	pêssego	0.00	3.06	6.11	0.00	0.00	0.00	0.00
	pêra	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	maçã	0.00	9.35	0.00	0.00	0.00	0.00	0.00

		Documentos						
		Pedrógrão Grande	Penacova	Penafiel	Penalva	Penamacor	Penedono	Penela
Culturas	**milho**	43.79	44.81	29.54	32.59	9.17	21.39	40.74
	batata	22.13	20.12	6.04	11.07	10.06	16.10	22.13
	vinha	8.15	9.17	9.17	15.28	6.11	7.13	10.18
	azeite	7.58	6.50	4.33	4.33	9.74	1.08	12.99
	arroz	2.39	1.19	2.39	2.39	1.19	0.00	2.39
	centeio	9.17	19.35	11.20	18.33	20.37	31.57	9.17
	trigo	7.00	15.00	4.00	16.00	23.00	12.00	22.00
	nabo	7.84	1.57	3.13	7.84	0.00	6.27	0.00
	feijão	23.57	17.42	11.27	11.27	7.17	7.17	17.42
	grão	11.73	1.17	1.17	3.52	2.35	0.00	15.24
	aveia	0.00	11.09	1.23	6.16	2.46	0.00	6.16
	oliveira	9.57	35.08	3.19	4.25	6.38	4.25	14.88
	amêndoa	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cevada	6.28	10.05	0.00	11.30	3.77	2.51	0.00
	chícharo	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cortiça	0.00	10.20	0.00	0.00	0.00	12.24	0.00
	noz	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	castanha	1.93	0.00	0.00	3.85	0.00	17.33	0.00
	avelã	0.00	0.00	0.00	11.39	0.00	0.00	0.00
	tremoço	2.95	5.90	1.47	0.00	0.00	1.47	0.00
	laranja	0.00	0.00	2.17	2.17	0.00	0.00	0.00
	limão	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mel	2.48	3.72	2.48	4.96	1.24	1.24	3.72
	aguardente	0.00	2.82	2.82	1.41	4.23	0.00	0.00
	pêssego	0.00	0.00	0.00	3.06	0.00	0.00	0.00
	pêra	32.09	0.00	0.00	6.42	0.00	0.00	0.00
	maçã	0.00	0.00	0.00	2.34	0.00	0.00	0.00

		Documentos						
		PesoDaRegua	Pinhel	Pombal	PonteDaBarca	PonteDeLima	Porto de Mós	Porto
Culturas	**milho**	25.46	15.28	77.40	86.57	43.79	67.22	17.31
	batata	9.06	26.16	35.21	29.18	23.14	14.09	15.09
	vinha	50.92	15.28	29.54	16.30	13.24	24.44	5.09
	azeite	7.58	10.83	14.07	11.91	4.33	21.65	0.00
	arroz	1.19	2.39	59.73	1.19	1.19	1.19	1.19
	centeio	17.31	39.72	6.11	45.83	21.39	1.02	2.04
	trigo	19.00	26.00	45.00	15.00	12.00	42.00	4.00
	nabo	3.13	3.13	1.57	4.70	3.13	0.00	4.70
	feijão	6.15	8.20	30.74	31.77	21.52	10.25	6.15
	grão	1.17	3.52	35.18	12.90	3.52	1.17	0.00
	aveia	9.86	2.46	6.16	0.00	11.09	9.86	2.46
	oliveira	13.82	2.13	31.89	9.57	7.44	19.13	1.06
	amêndoa	0.00	0.00	3.10	0.00	0.00	0.00	0.00
	cevada	12.56	0.00	2.51	0.00	2.51	10.05	3.77
	chícharo	0.00	0.00	3.90	0.00	0.00	0.00	0.00
	cortiça	0.00	28.55	0.00	22.43	0.00	0.00	0.00
	noz	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	castanha	0.00	7.70	0.00	19.25	1.93	0.00	0.00
	avelã	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	tremoço	7.37	1.47	1.47	14.75	0.00	8.85	1.47
	laranja	0.00	2.17	0.00	8.67	2.17	0.00	0.00
	limão	0.00	0.00	0.00	0.00	0.00	0.00	12.46
	mel	1.24	2.48	7.44	8.68	0.00	1.24	0.00
	aguardente	1.41	9.86	1.41	4.23	0.00	1.41	1.41
	pêssego	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	pêra	0.00	0.00	0.00	0.00	0.00	0.00	3.21
	maçã	0.00	2.34	0.00	7.02	0.00	0.00	0.00

		Documentos						
		Póvoa de Lanhoso	Póvoa de Varzim	Proença-a-Nova	Resende	Ribeira da Pena	Sabrosa	Sabugal
Culturas	**milho**	67.22	33.61	8.15	102.86	35.65	27.50	5.09
	batata	31.19	42.26	8.05	37.23	22.13	18.11	30.18
	vinha	23.42	5.09	0.00	48.89	6.11	21.39	8.15
	azeite	9.74	0.00	11.91	12.99	4.33	4.33	3.25
	arroz	1.19	3.58	1.19	1.19	1.19	1.19	4.78
	centeio	31.57	3.06	11.20	58.05	19.35	11.20	23.42
	trigo	13.00	9.00	9.00	31.00	2.00	6.00	19.00
	nabo	1.57	0.00	0.00	0.00	3.13	0.00	4.70
	feijão	43.04	9.22	3.07	32.79	9.22	17.42	5.12
	grão	4.69	0.00	2.35	14.07	1.17	1.17	1.17
	aveia	0.00	4.93	1.23	1.23	0.00	3.70	2.46
	oliveira	21.26	0.00	8.50	18.07	0.00	8.50	5.31
	amêndoa	0.00	0.00	0.00	0.00	0.00	0.00	3.10
	cevada	0.00	3.77	1.26	2.51	0.00	3.77	1.26
	chícharo	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cortiça	0.00	0.00	0.00	2.04	0.00	0.00	20.39
	noz	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	castanha	1.93	0.00	0.00	30.81	0.00	0.00	17.33
	avelã	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	tremoço	7.37	0.00	0.00	0.00	0.00	7.37	0.00
	laranja	4.34	0.00	4.34	0.00	6.50	0.00	0.00
	limão	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mel	1.24	2.48	4.96	2.48	0.00	1.24	6.20
	aguardente	0.00	2.82	1.41	4.23	0.00	1.41	4.23
	pêssego	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	pêra	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	maçã	2.34	0.00	0.00	2.34	0.00	0.00	0.00

		Documentos						
		Santa Comba Dão	Santa Maria da Feira	Santa Marta de Penaguiao	SantoTirso	São João da Madeira	São João da Pesqueira	São Pedro do Sul
Culturas	**milho**	52.96	67.22	17.31	69.26	50.92	16.30	60.09
	batata	23.14	34.21	3.02	33.20	29.18	15.09	12.07
	vinha	19.35	15.28	28.52	15.28	7.13	32.59	7.13
	azeite	10.83	3.25	1.08	5.41	3.25	6.50	9.74
	arroz	3.58	7.17	1.19	1.19	0.00	2.39	2.39
	centeio	12.22	11.20	9.17	22.41	7.13	28.52	10.18
	trigo	17.00	10.00	13.00	22.00	6.00	22.00	9.00
	nabo	4.70	1.57	6.27	6.27	0.00	4.70	3.13
	feijão	32.79	21.52	10.25	22.54	17.42	8.20	16.40
	grão	8.21	11.73	1.17	5.86	0.00	0.00	2.35
	aveia	4.93	6.16	2.46	8.63	16.02	1.23	2.46
	oliveira	10.63	11.69	7.44	1.06	7.44	12.75	13.82
	amêndoa	0.00	0.00	0.00	0.00	0.00	6.21	3.10
	cevada	6.28	8.79	3.77	10.05	10.05	0.00	0.00
	chícharo	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cortiça	0.00	2.04	0.00	0.00	0.00	0.00	2.04
	noz	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	castanha	0.00	0.00	0.00	0.00	0.00	0.00	1.93
	avelã	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	tremoço	5.90	7.37	4.42	14.75	1.47	1.47	1.47
	laranja	0.00	2.17	0.00	10.84	0.00	0.00	2.17
	limão	0.00	0.00	0.00	4.15	0.00	0.00	0.00
	mel	6.20	3.72	1.24	1.24	0.00	3.72	1.24
	aguardente	2.82	0.00	2.82	0.00	1.41	2.82	0.00
	pêssego	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	pêra	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	maçã	2.34	0.00	0.00	0.00	0.00	0.00	0.00

		Documentos						
		Sátão	Seia	Sernancelhe	Sertã	SeverDoVouga	Soure	Tabua
Culturas	**milho**	94.72	45.83	37.68	19.35	42.78	74.35	74.35
	batata	50.31	13.08	29.18	12.07	14.09	39.24	24.15
	vinha	23.42	10.18	13.24	3.06	10.18	46.85	19.35
	azeite	3.25	9.74	11.91	10.83	8.66	11.91	8.66
	arroz	2.39	0.00	1.19	1.19	1.19	51.37	1.19
	centeio	49.90	41.76	35.65	7.13	11.20	5.09	11.20
	trigo	27.00	5.00	18.00	14.00	3.00	42.00	11.00
	nabo	6.27	1.57	3.13	0.00	0.00	3.13	1.57
	feijão	19.47	17.42	10.25	7.17	36.89	39.96	25.62
	grão	12.90	3.52	5.86	1.17	2.35	7.04	11.73
	aveia	4.93	2.46	16.02	7.39	4.93	20.95	9.86
	oliveira	11.69	19.13	5.31	5.31	19.13	17.01	26.57
	amêndoa	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cevada	13.81	8.79	15.07	8.79	2.51	11.30	8.79
	chícharo	0.00	0.00	0.00	0.00	0.00	15.61	0.00
	cortiça	4.08	0.00	2.04	0.00	0.00	0.00	2.04
	noz	3.53	0.00	0.00	0.00	0.00	0.00	0.00
	castanha	13.48	3.85	25.03	0.00	0.00	0.00	0.00
	avelã	0.00	0.00	3.80	0.00	0.00	0.00	0.00
	tremoço	1.47	1.47	0.00	1.47	0.00	11.80	5.90
	laranja	0.00	0.00	0.00	0.00	15.18	0.00	0.00
	limão	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mel	3.72	1.24	0.00	8.68	0.00	1.24	6.20
	aguardente	1.41	11.27	1.41	8.45	2.82	5.63	1.41
	pêssego	0.00	0.00	0.00	0.00	3.06	0.00	0.00
	pêra	3.21	0.00	0.00	0.00	0.00	0.00	0.00
	maçã	16.37	0.00	2.34	0.00	4.68	0.00	2.34

		Documentos						
		Tabuaço	Tarouca	TerrasDeBouro	Tondela	TorreDeMoncorvo	Trancoso	Vagos
Culturas	**milho**	17.31	77.40	78.42	43.79	3.06	6.11	21.39
	batata	11.07	39.24	36.22	15.09	1.01	26.16	25.15
	vinha	21.39	19.35	19.35	21.39	8.15	7.13	5.09
	azeite	2.17	14.07	8.66	4.33	9.74	3.25	2.17
	arroz	2.39	2.39	0.00	1.19	0.00	1.19	23.89
	centeio	20.37	56.02	42.78	9.17	11.20	15.28	6.11
	trigo	12.00	21.00	4.00	8.00	18.00	7.00	16.00
	nabo	0.00	0.00	1.57	1.57	1.57	3.13	6.27
	feijão	6.15	24.59	54.31	22.54	1.02	5.12	15.37
	grão	0.00	3.52	4.69	17.59	0.00	0.00	7.04
	aveia	0.00	2.46	0.00	3.70	1.23	1.23	9.86
	oliveira	8.50	7.44	5.31	12.75	6.38	3.19	6.38
	amêndoa	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cevada	3.77	2.51	1.26	3.77	2.51	2.51	13.81
	chícharo	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cortiça	0.00	0.00	0.00	0.00	22.43	0.00	0.00
	noz	0.00	3.53	0.00	0.00	0.00	0.00	0.00
	castanha	0.00	3.85	0.00	0.00	21.18	0.00	0.00
	avelã	0.00	0.00	0.00	3.80	0.00	0.00	0.00
	tremoço	0.00	7.37	0.00	1.47	0.00	0.00	0.00
	laranja	0.00	0.00	2.17	2.17	0.00	0.00	8.67
	limão	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mel	3.72	3.72	4.96	1.24	0.00	9.92	0.00
	aguardente	1.41	0.00	2.82	1.41	2.82	2.82	0.00
	pêssego	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	pêra	0.00	3.21	0.00	0.00	0.00	0.00	0.00
	maçã	0.00	0.00	0.00	2.34	0.00	0.00	0.00

		Documentos						
		Vale de Cambra	Valença	Valongo	Valpaços	Viana do Castelo	Vieira do Minho	Vila de Rei
Culturas	**milho**	48.89	68.24	27.50	3.06	12.22	44.81	21.39
	batata	23.14	19.12	14.09	5.03	5.03	30.18	10.06
	vinha	15.28	9.17	4.07	11.20	3.06	14.26	7.13
	azeite	0.00	7.58	0.00	1.08	0.00	7.58	8.66
	arroz	0.00	1.19	0.00	0.00	0.00	2.39	1.19
	centeio	16.30	28.52	7.13	7.13	0.00	34.63	4.07
	trigo	12.00	29.00	2.00	17.00	6.00	5.00	10.00
	nabo	0.00	4.70	9.40	1.57	0.00	1.57	0.00
	feijão	19.47	19.47	6.15	5.12	4.10	27.67	4.10
	grão	15.24	9.38	15.24	0.00	0.00	7.04	2.35
	aveia	3.70	8.63	2.46	0.00	3.70	1.23	4.93
	oliveira	12.75	8.50	2.13	7.44	0.00	13.82	8.50
	amêndoa	0.00	3.10	0.00	0.00	0.00	0.00	3.10
	cevada	7.53	0.00	2.51	0.00	0.00	0.00	3.77
	chícharo	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cortiça	4.08	0.00	0.00	0.00	0.00	2.04	0.00
	noz	0.00	7.07	0.00	0.00	0.00	0.00	0.00
	castanha	1.93	0.00	0.00	0.00	0.00	1.93	0.00
	avelã	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	tremoço	4.42	0.00	1.47	0.00	0.00	0.00	4.42
	laranja	2.17	0.00	0.00	0.00	0.00	2.17	4.34
	limão	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mel	3.72	0.00	0.00	0.00	0.00	3.72	2.48
	aguardente	2.82	0.00	0.00	0.00	0.00	0.00	1.41
	pêssego	0.00	0.00	0.00	0.00	0.00	0.00	3.06
	pêra	0.00	0.00	0.00	0.00	0.00	0.00	3.21
	maçã	0.00	0.00	0.00	0.00	0.00	2.34	2.34

		Documentos						
		Vila do Conde	Vila Flor	Vila Nova de Cerveira	Vila Nova de Famalicão	Vila Nova de Foz Côa	Vila Nova de Gaia	Vila Nova de Paiva
Culturas	**milho**	71.29	2.04	22.41	74.35	11.20	33.61	68.24
	batata	30.18	8.05	10.06	49.30	9.06	16.10	30.18
	vinha	3.06	9.17	8.15	9.17	37.68	5.09	17.31
	azeite	3.25	8.66	1.08	6.50	5.41	2.17	2.17
	arroz	5.97	2.39	0.00	1.19	1.19	3.58	1.19
	centeio	20.37	12.22	13.24	29.54	23.42	6.11	51.94
	trigo	34.00	14.00	7.00	20.00	28.00	5.00	14.00
	nabo	10.97	1.57	0.00	3.13	0.00	0.00	0.00
	feijão	36.89	0.00	5.12	31.77	4.10	4.10	21.52
	grão	5.86	0.00	1.17	1.17	2.35	0.00	7.04
	aveia	20.95	0.00	2.46	12.32	2.46	0.00	0.00
	oliveira	0.00	5.31	1.06	17.01	8.50	8.50	6.38
	amêndoa	0.00	3.10	0.00	0.00	12.42	0.00	0.00
	cevada	15.07	0.00	2.51	11.30	18.84	3.77	3.77
	chícharo	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	cortiça	0.00	18.35	4.08	4.08	16.32	0.00	0.00
	noz	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	castanha	0.00	3.85	0.00	0.00	0.00	0.00	7.70
	avelã	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	tremoço	0.00	0.00	0.00	16.22	5.90	1.47	0.00
	laranja	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	limão	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	mel	6.20	1.24	2.48	3.72	4.96	2.48	3.72
	aguardente	2.82	1.41	1.41	0.00	1.41	1.41	0.00
	pêssego	0.00	0.00	3.06	0.00	0.00	0.00	0.00
	pêra	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	maçã	0.00	0.00	0.00	0.00	0.00	0.00	2.34

		Documentos				
		Vila Nova de Poiares	Vila Pouca de Aguiar	Vila Real	Vila Velha de Rodão	Vila Verde
Culturas	**milho**	44.81	42.78	36.66	21.39	40.74
	batata	21.13	52.32	26.16	9.06	17.10
	vinha	11.20	6.11	33.61	7.13	22.41
	azeite	14.07	0.00	6.50	11.91	11.91
	arroz	2.39	1.19	2.39	1.19	1.19
	centeio	11.20	30.55	21.39	4.07	15.28
	trigo	13.00	9.00	24.00	10.00	8.00
	nabo	0.00	3.13	0.00	0.00	4.70
	feijão	14.35	5.12	12.30	4.10	11.27
	grão	23.45	0.00	2.35	2.35	2.35
	aveia	12.32	1.23	3.70	4.93	2.46
	oliveira	7.44	2.13	2.13	15.94	10.63
	amêndoa	0.00	0.00	0.00	0.00	0.00
	cevada	5.02	1.26	5.02	3.77	0.00
	chícharo	0.00	0.00	0.00	0.00	0.00
	cortiça	2.04	0.00	0.00	0.00	32.63
	noz	0.00	0.00	0.00	0.00	0.00
	castanha	0.00	3.85	0.00	0.00	5.78
	avelã	0.00	0.00	0.00	0.00	0.00
	tremoço	7.37	0.00	10.32	4.42	8.85
	laranja	0.00	0.00	0.00	2.17	13.01
	limão	0.00	0.00	0.00	0.00	4.15
	mel	4.96	4.96	4.96	2.48	2.48
	aguardente	0.00	0.00	2.82	1.41	1.41
	pêssego	0.00	0.00	0.00	3.06	0.00
	pêra	0.00	0.00	0.00	0.00	3.21
	maçã	0.00	9.35	9.35	2.34	0.00

		Documentos			
		Vimioso	Vinhais	Viseu	Vouzela
Culturas	**milho**	7.13	5.09	64.16	78.42
	batata	10.06	17.10	35.21	17.10
	vinha	12.22	3.06	32.59	18.33
	azeite	14.07	0.00	10.83	11.91
	arroz	2.39	0.00	2.39	1.19
	centeio	38.70	17.31	19.35	38.70
	trigo	38.00	19.00	12.00	6.00
	nabo	6.27	0.00	12.54	3.13
	feijão	5.12	0.00	33.81	33.81
	grão	4.69	2.35	4.69	2.35
	aveia	1.23	0.00	2.46	1.23
	oliveira	8.50	2.13	5.31	11.69
	amêndoa	0.00	0.00	0.00	0.00
	cevada	2.51	0.00	12.56	3.77
	chícharo	0.00	0.00	0.00	0.00
	cortiça	42.83	0.00	0.00	0.00
	noz	3.53	0.00	0.00	0.00
	castanha	5.78	0.00	0.00	0.00
	avelã	0.00	0.00	15.19	0.00
	tremoço	1.47	0.00	8.85	0.00
	laranja	0.00	0.00	0.00	0.00
	limão	0.00	0.00	0.00	0.00
	mel	3.72	0.00	3.72	2.48
	aguardente	0.00	0.00	4.23	2.82
	pêssego	0.00	0.00	3.06	0.00
	pêra	0.00	0.00	0.00	0.00
	maçã	0.00	0.00	4.68	2.34