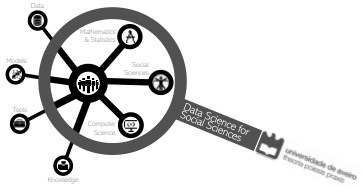

INTRODUÇÃO À CIÊNCIA DE DADOS

ANO 2023/2024

Perspetivas Globais sobre a Migração: Variações na Abordagem Científica entre os Países

Ana Beatriz Fernandes Pinho[1]

Beatriz de Almeida Gomes[2]

[1]*Mestrado em Ciência de Dados para Ciências Sociais*

[2]*Mestrado em Ciência de Dados para Ciências Sociais*

Resumo

Este artigo tem como objetivo analisar a abordagem da migração pela ciência de dados, explorando técnicas e consequentemente as conclusões obtidas. O projeto foi inicializado com uma recolha da Scopus com 454 artigos utilizando a pesquisa: ("immigration" OR "emigration" OR "migration") AND ("big data" OR "data processing" OR "data analysis") AND ("policy implications" OR "population dynamics"). A utilização da API da Scopus, e a utilização da programação de Python foi essencial para tornar os abstracts mais acessíveis. Estes artigos foram analisados por meio de técnicas avançadas de processamento textual, como Processamento de Linguagem Natural (NLP). Estas técnicas permitiram a identificação e representação das palavras-chave, servindo como base para os resultados obtidos. Para uma melhor visualização dos dados obtidos foi desenvolvida uma dashboard através da biblioteca Streamlit.

Palavras-chave: Migração, Ciência de Dados, Processamento de Linguagem Natuaral (NLP), Dashboard

1. Introdução

A ciência de dados tornou-se uma parte integral do nosso quotidiano, moldando e aprimorando diversas áreas das nossas vidas. A sua presença reflete-se numa variedade de setores, influenciando significativamente a forma como interagimos com o mundo digital e físico. A ciência de dados está na origem de muitas inovações que facilitam o nosso dia a dia. Desta forma, e segundo da IBM “What Is Data Science?”, a ciência de dados combina a estatística com a programação para extrair insights valiosos a partir de conjuntos de dados complexos.

A migração, um processo intrincado categorizado por espaço, legalidade, tempo e forma, apresenta diversas facetas, como migração rural-urbana e urbana-rural, além da distinção entre imigrantes legais e ilegais, migração de curto e longo prazo, voluntária e involuntária (*What Is Migration - Types, Causes & Impact*, n.d.).

O The World Migration Report 2020 destaca que, embora a maioria das migrações ocorra dentro dos países, o número de migrantes internacionais tem aumentado mais rapidamente do que previsto, com variações significativas entre as nações. A migração é um fenômeno intrinsecamente ligado à evolução

humana, sendo moldado por fatores geopolíticos, socioeconômicos e culturais.

Diante da complexidade contemporânea da migração, o projeto propõe uma abordagem inovadora, incorporando a ciência de dados para compreender as diversas perspetivas científicas adotadas pelos países. Isso visa identificar como tais perspetivas influenciam políticas, pesquisas e o movimento da população, contribuindo para o desenvolvimento de estratégias mais eficazes no enfrentamento dos desafios migratórios globais.

2. Metodologia

De forma a perceber como está a ser analisada a abordagem da migração pela ciência de dados foi conduzida uma abordagem bibliométrica e uma abordagem de análise de conteúdo de artigos que foram recentemente publicados pela comunidade científica.

Desta forma, a metodologia utilizada para a recolha de artigos foi dividida por 5 fases principais:

- Definição de uma questão de investigação;
- Recolha de Dados
- Redução dos Dados
- Análise e discussão de resultados;
- Visualização de resultados

Numa primeira fase do projeto, é crucial estabelecer uma questão de investigação clara e significativa. Com o intuito de compreender a abordagem da migração sob a perspetiva da ciência de dados, a pergunta de pesquisa proposta é a seguinte: “Como é que a pesquisa científica aborda a temática sobre migração entre diferentes países, e quais são os temas predominantes nessas abordagens?”

Na segunda fase do projeto, são estabelecidos os parâmetros específicos que guiarão a recolha de dados. Para a formulação da query ideal, foram necessárias diversas reformulações e tentativas até encontrar o conjunto de palavras-chave mais apropriado. Dentro do escopo deste projeto, para a extração de artigos da Scopus, foi adotada a seguinte query final: ('immigration' OR 'emigration' OR 'migration') AND ('big data' OR 'data processing' OR 'data analysis') AND ('policy implications' OR 'population dynamics'). Esta *query* revela uma abordagem abrangente, pois seleciona artigos que abordam não só os aspetos da migração, como também a dimensão da análise de dados, assim como as implicações políticas e as dinâmicas populacionais.

Assim, após a definição desses critérios, foi possível assegurar que os dados recolhidos sejam pertinentes e alinhados com os objetivos da pesquisa, permitindo, assim, uma análise mais precisa e significativa das perspetivas globais sobre a migração e as suas variações na abordagem científica entre os países.

Numa terceira fase foi fazer o pré processamento de dados. Esta limpeza foi subdividida em três grandes grupos: eliminação, segmentação e normalização.

Na primeira fase do pré-processamento de dados, o objetivo era remover todos os elementos desnecessários no texto, como pontuações e caracteres especiais. Em seguida, na segunda fase, procurámos segmentar o texto, convertendo todas as palavras para minúsculas para garantir consistência. Eliminámos as stopwords e palavras com baixa relevância informativa, além de aplicar o Steaming com o objetivo de reduzir as palavras às suas raízes para abranger diversas variações. Por fim, aplicámos a lematização para transformar as palavras nas suas formas canónicas.

Para representar os dados, implementámos a vectorização em gramas (por exemplo, unigramas, bigramas, trigramas). Além disso, conduzimos uma análise de sentimentos, atribuindo valores de 0 a 1, onde valores mais próximos de 1 indicam uma conotação mais positiva.

Numa quarta fase foi realizada uma análise e discussão dos resultados obtidos. Para analisar os artigos selecionados, foram adotadas diversas abordagens, integrando técnicas avançadas como a análise de clusters através do VOSviewer. Além da análise bibliométrica, que quantificou métricas como citações e fator de impacto, aplicou-se uma análise de conteúdo para compreensão qualitativa dos temas presentes nos artigos. A combinação destas abordagens proporcionou uma análise abrangente, explorando tanto as dimensões quantitativas quanto qualitativas da pesquisa sobre migração na ciência de dados.

Na última fase, desenvolvemos uma dashboard com o propósito de apresentar de forma clara todos os

resultados obtidos. Essa ferramenta proporciona ao utilizador uma compreensão autónoma. Desta forma, foi tido em atenção os seguintes parâmetros: Organização de análise de dados, apresentação de informação útil, conhecimento prático, aspeto estético e fácil compreensão e manipulação. Assim a dashboard tenta responder a algumas perguntas como Quem?, Onde?, Como? e Porquê?, relacionadas à produção dos resultados. Essa abordagem visa facilitar a interpretação e a utilização eficiente da ferramenta por parte do utilizador.

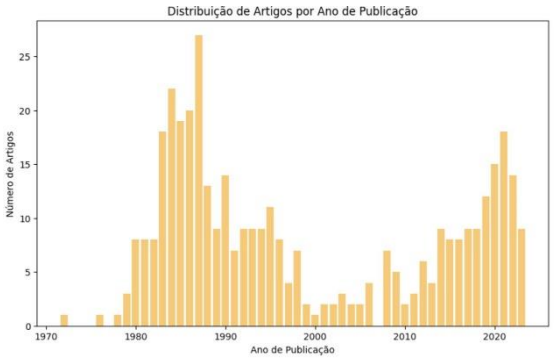

Figura 1 - Distribuição de Artigos por Ano de Publicação. Adaptado da Scopus

3. Resultados e Discussão

3.1. Análise Bibliométrica

A análise bibliométrica tem como objetivo realizar uma caracterização numérica abrangente dos artigos, incluindo informações como o ano de publicação, os autores envolvidos, a localização geográfica das pesquisas, as instituições de afiliação e os países de origem.

Assim, a análise bibliométrica proporciona uma visão detalhada das tendências temporais, identifica os principais contribuidores na área de estudo, destaca as regiões geográficas mais ativas e permite compreender as redes de colaboração entre instituições e países. A utilização desta ferramenta da Scopus fornece uma base sólida para a investigação, permitindo uma avaliação abrangente da produção científica relacionada à migração e ciência de dados.

Para realizar a análise bibliométrica, foram utilizadas duas ferramentas principais: a ferramenta de análise de resultados disponibilizada pela Scopus, que proporcionou informações mais simples e diretas, e o VOSviewer, que permitiu agrupar os documentos retornados numa rede de ligações e explorar as relações entre eles.

Distribuição de artigos pelo ano de publicação revela dois picos notáveis, conforme evidenciado na figura 1, relativos ao número de publicações científicas sobre o tema em questão. O primeiro pico remonta à Guerra Fria, um período que teve impacto global e impulsionou movimentações populacionais. O segundo pico, ocorrido nos anos de 2020, inicialmente associado à pandemia, surpreendentemente está relacionado à crise de 2008, como indicam os artigos publicados nesse período.

A análise dos domínios revela informações valiosas sobre a distribuição das publicações relacionadas ao tema. A Figura 2 destaca que as ciências sociais lideram em termos de número de publicações, representando significativos 27,4% do total. Em segundo lugar, observamos uma forte presença na área da Medicina, abrangendo aproximadamente 24,1% das publicações. Em terceiro lugar, a categoria de 'outras ciências' contribui de maneira considerável, representando 12,5% do panorama geral.

Esta distribuição sugere uma relevância particular do tema nas ciências sociais e na Medicina, indicando um interesse e investimento substancial nessas áreas. A presença significativa em 'outras ciências' sugere uma diversidade de abordagens e contribuições de diferentes disciplinas para o tema em questão. Esta análise dos domínios fornece insights valiosos para compreender a variedade de perspetivas e abordagens que convergem para o estudo em questão.

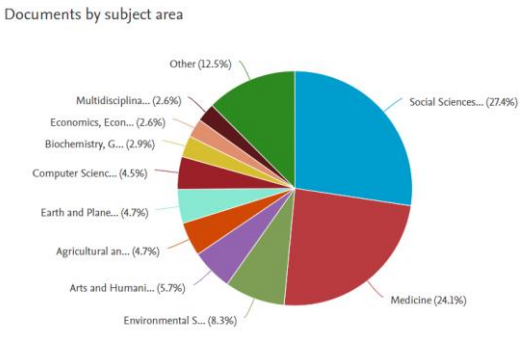

Figura 2 - Documentos por áreas científicas. Adaptado da Scopus

Agora **analisando os títulos** com recurso ao VOSviewer, é possível observar que a identificação das palavras mais frequentes e que possuem maior importância para a análise. Conforme evidenciado na Figura 3, os termos mais salientes são 'China', 'análise', 'migração' e 'estudo'.

A presença recorrente da palavra 'China' sugere uma ênfase significativa nos contextos relacionados a este país, indicando possíveis conexões ou focos temáticos específicos. O termo 'análise' evidencia a natureza analítica das abordagens adotadas nos artigos, apontando para uma metodologia robusta na avaliação dos temas em questão. A palavra 'migração' indica um interesse particular nas dinâmicas de migração, enquanto 'estudo' destaca a natureza investigativa dos trabalhos.

Estes resultados sugerem uma forte associação entre os temas de China, análise, migração e estudo nas publicações analisadas. Estas palavras-chave fundamentais fornecem pistas importantes sobre os enfoques e conteúdos predominantes nos títulos dos artigos, contribuindo assim para uma compreensão mais aprofundada do escopo e do foco dessas pesquisas.

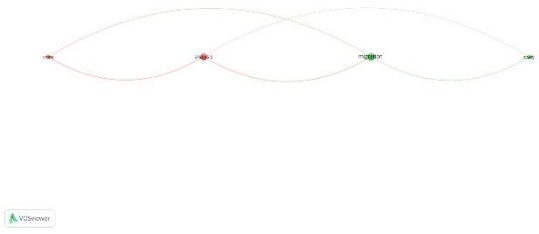

Figura 3 - Análise do Título. Adaptado pelo VOSviewer

A figura 4 é referente à **análise do abstract** e a figura 5 apresenta uma análise dos **títulos juntamente com os resumos (abstract)**, respetivamente. Notavelmente, ambas as figuras revelam a presença de três clusters distintos que denotam padrões semelhantes nos temas abordados. No primeiro cluster, representado pela cor azul, destacam-se termos como 'casamento', 'educação', 'área urbana', 'duração', 'idade', 'mulher', 'criança', 'sexo', 'morte', 'projeção', 'masculino', 'declínio', 'crescimento', 'domicílio', 'migração urbana', 'área rural', 'aumento' e outros. Este cluster sugere uma ênfase nos aspetos demográficos, casamento, educação e migração.

No segundo cluster, identificado pela cor verde, observam-se termos como 'renda', 'janeiro', 'China', 'cidade', 'risco', 'efeito', 'controle', 'urbanização', 'impacto', 'papel', 'influência', 'adição', 'mobilidade', 'modelo', 'big data', 'localização', 'mundo', 'fluxo', 'base', 'processo' e 'abordagem'. Esses termos indicam uma abordagem mais centrada em aspetos económicos, risco, urbanização e modelagem, com destaque para big data.

No terceiro cluster, representado pela cor vermelha, surgem termos como 'residência', 'local', 'origem', 'destino', 'questão', 'do autor', 'autor', 'estimativa', 'México', 'migração internacional', 'emigração', 'dados censos', 'exemplo', 'uso', 'aplicação', 'imigrante', 'tipo' e 'contexto'. Este cluster aponta para uma análise mais centrada em questões relacionadas à residência, origem e destino, destacando a aplicação de dados dos censos e o contexto de migração internacional.

A consistência entre os clusters nas duas figuras sugere uma convergência nos temas abordados nos resumos e nos títulos e resumos combinados, reforçando a importância desses tópicos na pesquisa em questão.

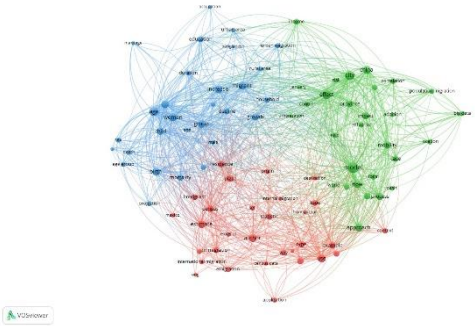

Figura 4 - Análise do Abstract. Adaptado pelo VOSviewer

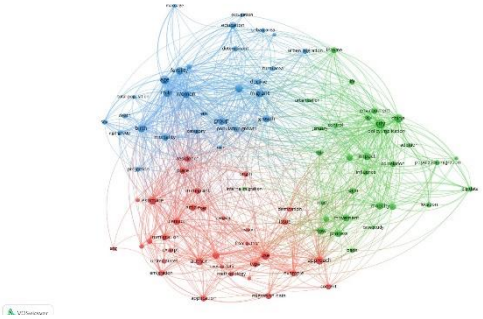

Figura 5 - Análise do Título e do Abstract. Adaptado pelo VOSviewer

3.2. Análise de Conteúdo

Foi elaborado um notebook para conduzir uma análise abrangente do conteúdo dos artigos científicos, dividido em quatro fases distintas. A primeira fase engloba o 'Pré-processamento de dados', onde são aplicadas técnicas para limpar, normalizar e organizar os dados, assegurando que estes estejam preparados para análises subsequentes.

Na segunda fase, abordamos a 'Frequência de termos', utilizando métodos que permitem identificar os termos mais frequentes nos textos dos artigos. Este passo proporciona uma visão inicial dos temas predominantes nas publicações, destacando palavras-chave relevantes.

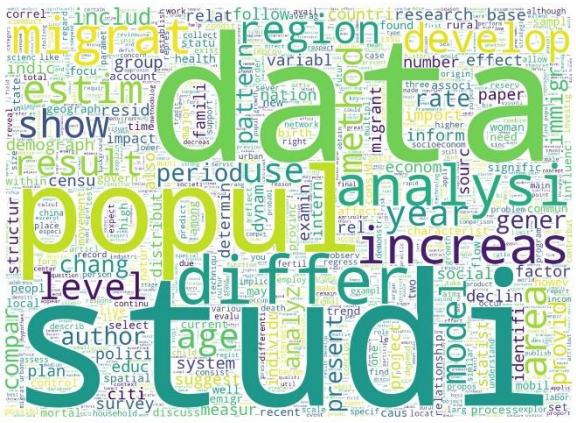

Figura 6 - Termos mais Comuns no Abstract (sem as stopwords)

A terceira etapa envolve a 'Representação numérica de texto', onde exploramos n-gramas e bigramas para capturar relações semânticas e contextuais entre as palavras nos artigos. Esta abordagem contribui para uma compreensão mais aprofundada da estrutura e do significado dos textos.

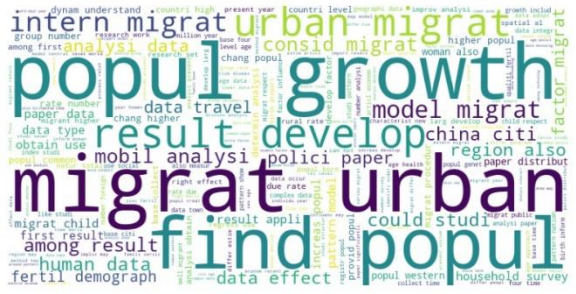

Figura 7 - Bigramas

Por último, a quarta fase consiste na aplicação de 'Topic Modelling' utilizando o algoritmo Latent Dirichlet Allocation (LDA). Assi, deu-se início à análise textual, utilizando a técnica de "topic modelling". Segundo Lande et al., (2023), o "topic modelling" é uma disciplina no âmbito do processamento de linguagem natural (NLP) que tem como objetivo detetar, analisar e classificar "tópicos" num corpus, identificando padrões e temas comuns. As abordagens tradicionais nesse campo incluem a Análise Semântica Latente (LSA) e a Alocação Latente de Dirichlet (LDA).

No decorrer deste projeto, foi utilizada a abordagem LDA. De acordo com Lande et al., (2023), esta abordagem assume que os documentos são combinações de vários tópicos, sendo cada tópico uma mistura de palavras com uma determinada pontuação de probabilidade. Assim, é definido um conjunto de tópicos. Para determinar o número ideal de tópicos que melhor explica os documentos em análise, utilizamos um modelo de coerência, avaliando a coerência de cada tópico para identificar o número mais apropriado. Concluímos que 8 tópicos era o número ótimo de tópicos segundo este método. Cada tópico conta com um conjunto de palavras-chave que são comparadas com as palavras-chave de cada documento.

Desta forma, foi possível obter 8 tópicos:

```
[(0,
'0.033*"population" + 0.013*"data" + 0.009*"migration" + 0.006*"model" + '0.005*"populations" + 0.005*"analysis" + 0.005*"growth" + '0.005*"development" + 0.004*"china" + 0.004*"based"'),

(1,
'0.010*"population" + 0.008*"data" + 0.007*"age" + 0.007*"migration" + '0.006*"used" + 0.005*"distribution" + 0.005*"areas" + 0.005*"water" + '0.005*"number" + 0.005*"also"'),

(2,
'0.023*"data" + 0.018*"migration" + 0.005*"time" + 0.004*"information" + "0.004*"research" + 0.004*"population" + 0.004*"model" + 0.004*"year" + "0.004*"method" + 0.003*"study"'),
```

```
    (3,
'0.022*"urban" + 0.016*"migration" +
0.013*"population" + 0.012*"rural" +
"0.010*"data" + 0.009*"migrants" +
0.008*"cities" + 0.006*"level" +
"0.005*"economic" + 0.005*"income"'),

    (4,
'0.018*"population" + 0.011*"migration" +
0.011*"age" + 0.011*"data" + "0.009*"urban" +
0.008*"migrants" + 0.008*"years" +
0.008*"fertility" + "0.006*"children" +
0.006*"higher"'),

    (5,
'0.011*"data" + 0.008*"migration" +
0.006*"mobility" + 0.006*"level" +
"0.006*"population" + 0.005*"cities" +
0.005*"policy" + 0.005*"analysis" +
"0.005*"growth" + 0.004*"health"'),

    (6,
'0.013*"migration" + 0.010*"population" +
0.010*"women" + 0.007*"data" +
"0.007*"available" + 0.006*"abstract" +
0.006*"urban" + 0.006*"survey" +
"0.005*"study" + 0.005*"migrants"'),

    (7,
'0.017*"data" + 0.014*"migration" +
0.010*"model" + 0.009*"spatial" +
"0.008*"population" + 0.006*"analysis" +
0.005*"fertility" + 0.005*"using" +
"0.005*"level" + 0.005*"results"')]
```

Desta forma, é possível analisar cada tópico.

Tópico 0: Demografia e Desenvolvimento (crescimento populacional)

Neste tópico, as palavras-chave sugerem uma abordagem centrada na demografia e no desenvolvimento, com ênfase no crescimento populacional. Questões relacionadas ao tamanho e dinâmica da população, juntamente com o seu impacto no desenvolvimento, podem ser temas importantes de salientar.

Tópico 1: Distribuição e Idade da População

O segundo tópico destaca a distribuição geográfica da população e a sua estrutura etária. A análise parece concentrar-se em como a idade influencia a distribuição populacional, com menções a dados relacionados à idade e migração.

Tópico 2: Análise de Dados e Migração ao Longo do Tempo

Este tópico sugere uma abordagem mais orientada para a análise de dados, com foco específico na migração ao longo do tempo. A utilização de métodos analíticos para compreender padrões temporais na migração pode ser um tema central.

Tópico 3: Migração Urbana e Rural

O quarto tópico destaca a dicotomia entre migração urbana e rural. Questões relacionadas ao movimento de populações entre áreas urbanas e rurais, bem como o seu impacto em níveis económicos e de renda, podem ser exploradas neste contexto.

Tópico 4: Demografia, Migração e Idade

Este tópico aborda a interseção entre demografia, migração e idade. A relação entre idade e migração, juntamente com os seus efeitos na estrutura populacional, pode ser um foco central de investigação.

Tópico 5: Mobilidade, Política e Saúde

O quinto tópico sugere uma abordagem mais ampla, abrangendo mobilidade populacional, políticas relacionadas e aspetos de saúde. Pode incluir análises sobre como a mobilidade afeta políticas públicas e saúde populacional.

Tópico 6: Migração de Mulheres

Neste tópico, a atenção parece voltar-se especificamente para a migração de mulheres. Questões relacionadas à participação das mulheres nos movimentos migratórios e o seu impacto social podem ser temas relevantes.

Tópico 7: representação aeroespacial e Análise da Fertilidade

O último tópico sugere uma abordagem espacial e destaca a análise da fertilidade. O uso de simulações espaciais pode ser central, juntamente com a análise dos fatores que influenciam a fertilidade populacional.

Estes tópicos proporcionam uma visão abrangente dos temas explorados nos artigos científicos, abrindo espaço para investigações mais detalhadas em cada uma destas áreas específicas.

Posteriormente, com o objetivo de evidenciar a neutralidade dos artigos, realizou-se uma Análise de Sentimentos, revelando que os documentos analisados apresentam predominantemente sentimentos neutros, indicando uma abordagem informativa e imparcial nos textos. A pontuação composta (compound) é próxima de zero para todos os documentos, sugerindo uma ausência geral de polaridade nas expressões. Contudo, é importante notar que o documento 9 possui uma pontuação negativa significativa (-0.296), indicando um leve tom negativo naquele artigo em específico.

Desta forma, o notebook abrange desde a preparação inicial dos dados até a extração de tópicos latentes, proporcionando uma análise aprofundada e abrangente do conteúdo dos artigos científicos em questão.

3.3. Dashboard

A dashboard foi desenvolvida utilizando a aplicação Streamlit, seguindo o princípio de organizar os dados de forma eficaz para destacar as informações mais relevantes e cruciais. Antes de criar a dashboard, foram cuidadosamente considerados diversos parâmetros, incluindo a definição do público-alvo. Dado que o tema central é a migração, abrangendo uma ampla gama de interesses, a decisão foi orientar a dashboard para a população em geral.

O aspeto estético também foi levado em consideração, procurando não só funcionalidade, mas também atratividade visual. O objetivo foi criar uma interface de fácil compreensão e manipulação, promovendo a simplicidade para utilizadores com diferentes níveis.

A dashboard está dividida em duas partes distintas: a primeira destina-se à visualização e análise específica de um único país, proporcionando uma análise mais aprofundada e detalhada. Já a segunda parte permite a comparação direta entre dois países, oferecendo uma perspetiva comparativa valiosa.

Os temas abordados incluem os anos de publicação, o resumo (abstract), as palavras-chave e os tópicos, assegurando uma análise abrangente e informativa.

4. Reflexões Finais

A investigação teve início com a pergunta de investigação: "Como é que a pesquisa científica aborda a temática sobre migração entre diferentes países, e quais são os temas predominantes nessas abordagens?"

Após uma análise abrangente e interpretação dos dados, tornou-se evidente que nos países europeus, a temática genética surge consistentemente. Essa associação torna-se clara ao examinar o tópico 7, que trata da análise da fertilidade populacional, revelando que os estudos sobre migração não se limitam apenas a esse fenómeno, mas exploram também as interligações com doenças, fertilidade e genética. Esta ênfase na genética explica o impacto significativo da área da medicina nos artigos publicados.

Observou-se que cada país aborda temas distintos, embora países geograficamente próximos apresentem termos semelhantes. As grandes tendências identificadas foram: a genética (na Europa e América do Sul), migração urbana e rural (na Ásia), distribuição e idade da população (na Oceânia) e dinâmicas populacionais (na América do Norte).

Estas descobertas destacam a diversidade de abordagens em relação à migração em diferentes regiões do mundo, enfatizando a importância de considerar fatores locais e regionais ao estudar este fenómeno complexo.

5. Bibliografia

Lande, J., Pillay, A., & Chandra, R. (2023). *Deep learning for COVID-19 topic modelling via Twitter: Alpha, Delta and Omicron*. http://arxiv.org/abs/2303.00135

The World Migration Report 2020. (n.d.). Retrieved January 8, 2024, from https://worldmigrationreport.iom.int/wmr-2020-interactive/

What is Data Science? | IBM. (n.d.). Retrieved January 8, 2024, from https://www.ibm.com/topics/data-science

What is Migration - Types, Causes & Impact. (n.d.). Retrieved January 8, 2024, from https://www.geeksforgeeks.org/what-is-migration-types-causes/