TP3 - AEDS

Vinícius Mello

Julho 2018

1 Introdução

O processo de globalização iniciado no começo do século vem moldando tanto a economia dos países em geral quanto a forma como produzimos bens de consumo. Nesse novo ambiente os países passam fazer parte de uma complexa rede de comercio internacional de onde surgem diferente estratégias de exportação. O presente trabalho tem como objetivo a análise de forma conjunta dos perfis de exportações dos países e como esses se agrupam.

Nesse estudo, cada país é caracterizado pela suas exportações em 21 classes de produtos. A classificação para produtos utilizado foi baseada na *Harmonized Commodity Description and Coding Systems* de 2007, e ao invés de utilizar os valores monetários de exportações foi utilizado um índice conhecido como *Revealed Comparive Advantage* (RCA) que representa a vatagem relativa de um pais no comercio de uma determinada classe de bens. A escolha desse índice foi feita por esse representar melhor a estratégia de exportação adotada por um país independente do seu porte. No decorrer do estudo os valores de RCA de cada classe serão considerados as features que descrevem um indivíduo (país) na análise feita.

Foram analisados os dados de RCA de cada país entre os anos de 2008 e 2014, apesar do foco da análise ser no mais recente. O objetivo da análise temporal e verificar se as estruturas encontradas se mantem ao longo dos anos ou se existe muita mudança.

2 Metodologia

O foco desse estudo é fazer um análise dos perfis de exportação de diferentes países e buscar uma estrutura natural de agrupamento de países. Para isso foram utilizados dados sobre exportação total em dólares de em 21 diferentes categorias de produtos, de acordo com a classificação HS de 2007. Essa classificação tem como objetivo proporcionar aos países participantes uma forma de classificar as suas trocas comerciais usando um padrão comum.

A classificação HS possui diferentes níveis que chegando em até aproximadamente cinto mil classes diferentes. Nesse estudo foi futilizado apenas o nível mais alto com 21 classes para facilitar o agrupamento. Um número muito grande de features iria tornar difícil o agrupamento dado que o número de instâncias (países) não é suficientemente grande. Os dados foram obtidos através do site https://atlas.media.mit.edu que reune informações de diversas fontes. Os dados analisados nesse estudo são referentes ao anos de 2008 a 2014.

Como o objetivo do trabalho é analisar a perfil de exportações dos países optou-se pela utilização do RCA (Revealed Comparive Advantage) ao invés dos valores financeiros de

exportação. Esse índice mede a vantagem relativa de um país no comercio de uma classe de produtos e tem a seguinte formula:

$$RCA_{pc} = \frac{E_{pc} / \sum_{c' \in C} E_{pc'}}{\sum_{p' \in P} E_{p'c} / \sum_{p' \in P} \sum_{c' \in C} E_{p'c'}}$$

Onde E_{pc} é a exportação do país p na classe c. Nesse índice o numerador é a proporção de uma determinada classe nas exportações de um país, e o denominador a proporção dessa classe nas exportações mundiais. Dessa forma esse índice será maior que um caso um país tenha uma participação maior que sua "cota justa" nas exportações da classe, e menor caso o contrário.

Por último, os valores de RCA obtidos passaram por uma transformação conhecida como normalização minmax, em que os valores são normalizados para ficarem entre zero e um. Essa transformação é necessária para algumas das técnicas utilizadas nesse estudo impedindo que a análise seja dominada por outliers.

Por ser de natureza exploratória, o presente trabalho fez uso de ferramentas de análise de uma forma sequencial, de forma que muitas das técnicas foram aplicadas sobre demanda de acordo com os insights obtidos anteriormente. Apesar disso, essa análise foi motivado por uma hipotese inicial de que deve haver um agrupamento natural dos países de acordo com um número reduzido de estratégias de exportação. Além disso, é de se esperar que esses grupos se mantenham relativamente constantes no curo prazo e que a mudança na dinâmica desses seja gradual. Como veremos nos resultados obtidos, apesar de que podemos identificar estrutura nos dados analisados, a variação anual é muito grande, e muitos grupos se formam e desaparecem de um ano para o outro. No restante dessa secção será discutidas as técnicas utilizadas nesse estudo, e na seção seguinte serão apresentados os resultados obtidos.

2.1 Análise de Componentes Principais

A análise de Componentes Principais (PCA) é uma técnica estatística muito utilizada em análise exploratória e redução de dimensionalidade. O método consiste em fazer uma transformação ortogonal nas features (normalmente correlacionadas) de um conjunto de observações de forma a representar esses em um novo espaço de variáveis não correlacionadas chamadas de Componentes. O número de componentes gerados pela técnica é igual ao número de features originais (caso existam mais observações que features) e essas são construídas de forma que a primeira componente explica a maior parte da variabilidade das observações. O intuito nesse trabalho da utilização dessa técnica é entender quais as classes (features) mais importantes para explicar a variabilidade dos dados.

2.2 t-Distributed Stochastic Neighbor Embedding

Também conhecido como t-SNE, essa técnica de Aprendizado de Máquina também é muito utilizado para redução de dimensionalidade, porém essa redução tem como principal objetivo a visualização, normalmente em duas dimensões, de dados com alto dimensionalidade. Esse mapeamento para coordenadas em duas dimensões é usado nesse estudo para possibilitar a utilização de gráficos e visualização dos grupos formados pelos demais métodos.

O método consiste em duas etapas, primeiro é calculado um distribuição de probabilidade sobre as observações em alta dimensionalidade. Depois uma segunda distribuição

é feita na dimensionalidade resultante de forma a minimizar a divergência de *Kullback–Leibler* das duas distribuições em relação a localização dos pontos no mapeamento.

2.3 Agrupamento Hierárquico

O primeiro método de agrupamento utilizado nesse estudo é o agrupamento hierárquico, que observações em grupos de com base em uma medida de distância definida. O funcionamento do algorítimo é simples, todas as observações são inicializadas em um grupo contendo apenas um elemento, e em cada passo os dois grupos mais próximos são unidos em um grupo contendo todos os seus elementos. Inicialmente o algorítimo finaliza apenas quando todos as observações estão em um grande grupo, e no processo um gráfico chamado dendrograma é formado mostrando as uniões feitas nos diferentes passos e a distância entre os grupos unidos. Com base nesse dendrograma e em métricas de qualidade o usuário deve determinar um ponto de parada, ou seja, um número de grupos formados.

Apesar de simples, o agrupamento hierárquico possue uma grande vantagem que é de possibilitar a visualização de suas iterações, que pode ajudar no entendimento da estrutura dos dados. Nesse trabalho a medida de distância utilizada foi a distância euclidiana.

2.4 K-Means

K-means é um método de agrupamento muito conhecido que, deferente do hierárquico deve ser inicializado já com o número de grupos resultantes. O grupos são inicializados normalmente de forma aleatória, e a partir de então começa um processo iterativo onde as observações são alocadas no grupo mais próximo com base em uma medida de distância, e os centros dos grupos são recalculados com base na centroide de seus componentes. Por ser dependente de um inicialização aleatória, o algorítimo pode resultar em agrupamentos diferentes para os mesmos dados. Por conta disso o ideal é que sejam feitas várias tentativas de agrupamento mantendo a que obtiver os melhores resultados.

Nesse estudo o método de K-meas é inicializado cinquenta vezes para garantir o melhor resultado possível. Assim como no método hierárquico a medida de distância usada foi a distância euclidiana, e vários valores para o número final de grupos são testados.

3 Resultados Computacionais

Nessa seção serão discutidos os resultados obtidos na aplicação das análises propostas. Começando pela análise exploratória inicial dos dados e em seguida os métodos de agrupamento. Como será explicado a seguir, a hipótese inicial de que deve existir um agrupamento natural que se mantem relativamente constante ao longo dos anos não foi confirmada, ao invez disso observamos uma estrutura geral que se mantem, mas agrupamentos que formam e somem de um ano para o outros, cateterizando um dinamismo maior do que o esperado.

3.1 Análise Exploratória

Com intuito de entender a estrutura geral dos dados de exportações dos países foi feita uma análise estatística básica e implementação de algumas técnicas para representação dos dados. Primeiramente foi calculado indicadores estatísticos básicos como média, desvio

padrão e os quatis das features, buscando identificar aquelas que serão de maior importância na análise. Na imagem 1 podemos ver esses resultados dessa análise para o ano de 2014 em um boxplot para cada feature.

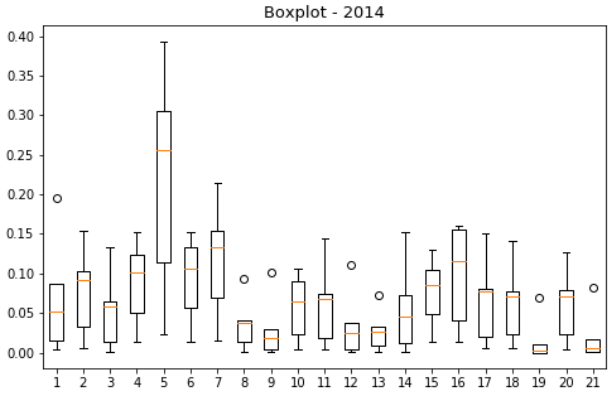


Figure 1: Boxplot das features

Esse gráfico deixa evidente que a variabilidade do quinta classe referente a *Mineral Products* tem uma variabilidade muito maior que as demais, além disso sua média também é mais elevada. Isso pode ser explicado pelo fato de que petróleo e seus derivados que são extremamente importantes para a economia atual se encontram nessa classe. No Anexo A temos os boxplots para cada ano, e podemos ver que essa configuração se mantem.

Para explorar mais a estrutura dos dados foi usado o método PCA con intuito de olhar para as primeiras componentes resultantes resultantes e observar quais features tem maior peso nessas. Primeiro, é necessário analisar a variância explicada para cada componente. Isso está presente na figura 2 que mostra a proporção da variância explicada

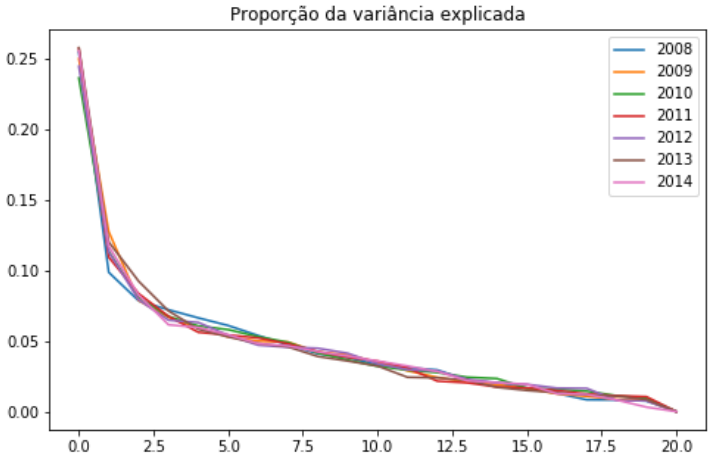


Figure 2: Análise PCA

Pelo gráfico podemos ver que precisamos de um grande número de componentes para explicar a variabilidade dos dados. Isso indica que a análise não está sendo dominada por nenhum fator. De fato para explicar aproximadamente setenta e cinco porcento da variabilidade em todos os anos precisamos de nove componentes. Como vemos na figura 3

que mostra os pesos das features em cada uma das seis primeiras componentes, só existe concordância do PCA entre os diferentes anos, nas primeiras componentes.

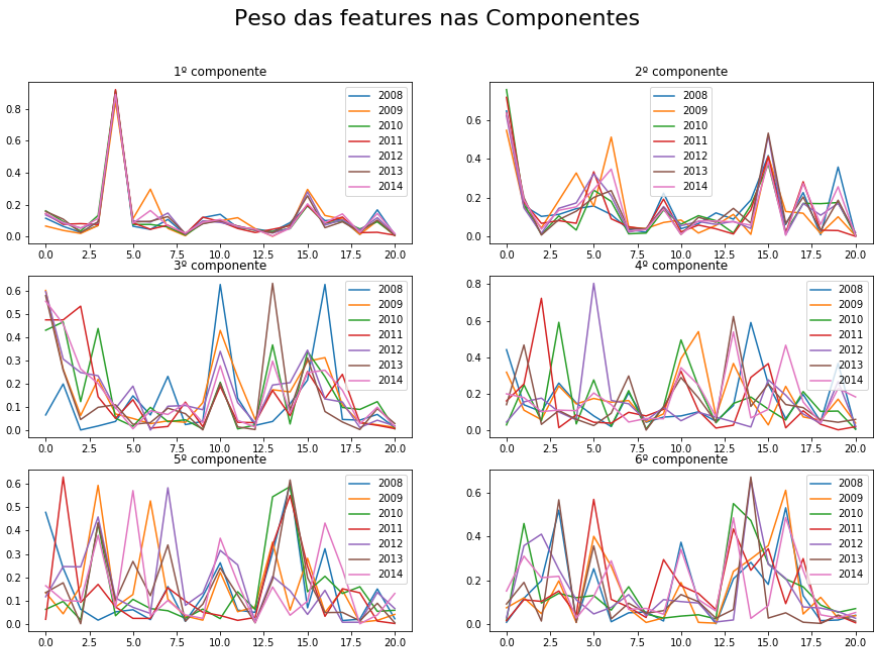


Figure 3: componentes do PCA

Essa análise começa a dar os indícios que os a estrutura dos dados tem dinâmica bem variada ao longo dos anos.

Para finalizar a exploração dos dados foi feito um mapeamento t-SNE *t-Distributed Stochastic Neighbor Embedding* para que os países possam ser observados graficamente em duas dimensões. Para tentar manter a posição de um país ao longo dos anos relativamente estável, o modelo foi aplicado em todos os dados (todos os anos), mas nesse estudo representaremos cada ano em um gráfico diferente. Na figura 4 temos o mapeamento da posição dos países em 2014 para demonstrar os resultados da técnica e no Anexo B são apresentados todos os gráficos.

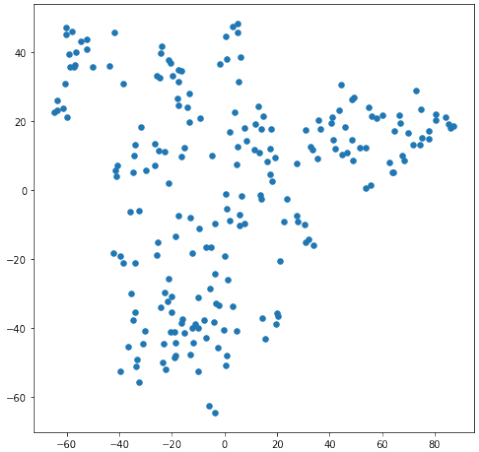


Figure 4: t-SNE

Esse mapeamento feito através do t-SNE será utilizado para ajudar a visualização dos gráficos nos modelos de agrupamentos a discutidos a seguir.

3.2 Agrupamento

Uma hipótese inicial que motivou a elaboração desse estudo, foi de que deve haver um agrupamento dos países relacionado ao perfil de exportação que tenha modificações graduais ao longo dos anos. Como primeiro passo na avaliação dese hipótese foi feito um agrupamento hierárquico nos dados. Como não sabemos a priori o número de grupos a serem formados, foi utilizado a métrica de avaliação de silhueta que é uma medida de quão similar é um elemento ao seu grupo quando comparado com os demais grupos. A tabela 3.2 mostra o número de grupos que maximiza essa medida para cada ano, e podemos perceber que o número de grupos formados variou muito de um ano para o outro, o que indica uma dinâmica pouco estável. Na figura 5 é representado o dendrograma, o mapeamento t-SNE e a o valor médio de em cada feature dos agrupamentos de 2014.

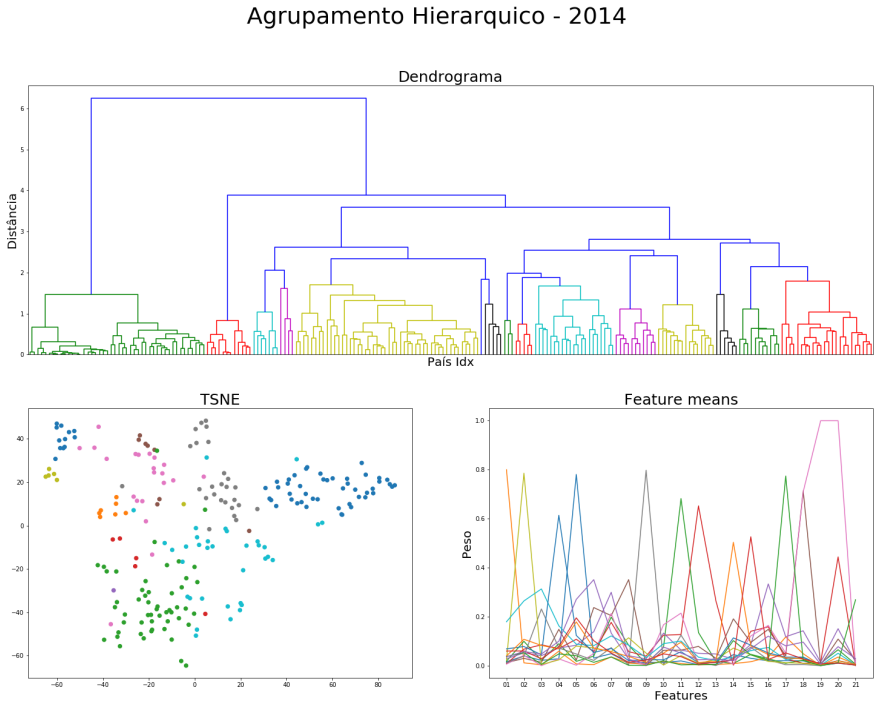


Figure 5: Agrupamento Hierárquico - 2014

Pelo dendrograma e gráfico do mapeamento t-SNE percebemos que o agrupamento consegue encontrar um estrutura nos dados. O gráfico das centroides mostra que cada grupo parece dominar algumas features, o que faz sentido já que os países iram buscar estratégias de diferenciação para competir no comercio internacional. Apesar disso, o agrupamento falha em construir grupos constantes no tempo. No Anexo C temos são apresentados os gráficos para todos os anos.

A aplicação do k-means nos dados foi feita de forma semelhante ao agrupamento hierárquico. Na tabela 3.2 temo o número de grupos que minimizou a métrica de silhueta, e quando comparado com o método de agrupamento hierárquico podemos perceber que

apesar de resultados semelhantes o k-means tende a resultar em um número menor de grupos. Na figura 6 podemos observar o resultado da aplicação do método para o ano de 2014.

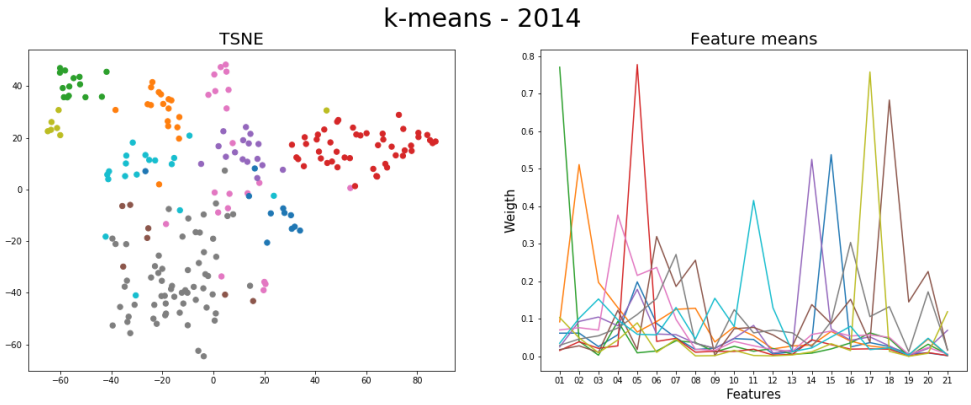


Figure 6: K-means

	Hierárquico		K-means	
Ano	Silhueta	num_clusters	Silhueta	num_clusters
2008	0.312835	15.0	0.290129	9.0
2009	0.278655	16.0	0.280519	9.0
2010	0.233286	10.0	0.267113	9.0
2011	0.315076	17.0	0.296935	13.0
2012	0.293350	14.0	0.290140	10.0
2013	0.309512	11.0	0.308026	14.0
2014	0.271725	15.0	0.286965	10.0

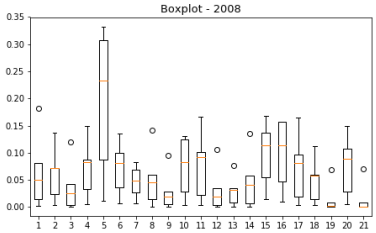
4 Conclusão e Trabalhos Futuros

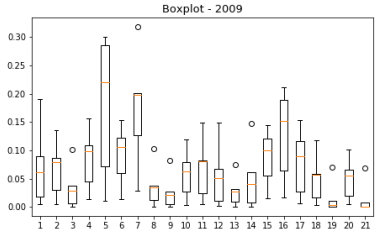
No presente trabalho, diversos métodos de aprendizado não supervisionado foram aplicados a um conjunto de dados de exportações de países com base em uma classificação internacional de produtos. O objetivo foi identificar agrupamentos naturais nos dados que pudessem explicar a dinâmica do comercio internacional. Como resultados obtivemos modelos capazes de agrupar anualmente os países de forma satisfatória o que pode auxiliar no entendimento da dinâmica do comercio internacional. Por outro lado os agrupamentos e estruturas encontrados sofrem grande alteração ao longo dos anos, e muitos grupos formam e se desfazem, dificultando uma análise temporal. Duas possíveis explicações foram elaboradas para esse resultado. A primeira é de que a dinamicidade do comercio internacional deve ser analisada em intervalos menores de tempo, e em conjunto com especialistas para que seja possível compreender a dinâmica temporal e a evolução dos grupos. A outra hipótese é de que apesar de conveniente para agrupar os produtos a classificação HS de 2007 utilizada não é ideal para entender dinâmicas comerciais.

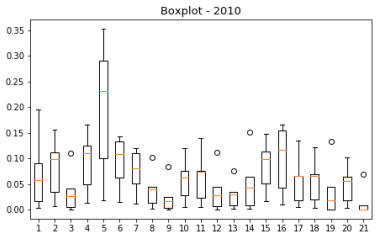
Uma proposta para trabalhos futuros, é a da utilização de uma classificação de nível mais baixo da HS de produtos seja utilizada, e um primeira fase de agrupamento das features, e redução de dimensionalidade seja utilizada, para em primeiro lugar verificar se

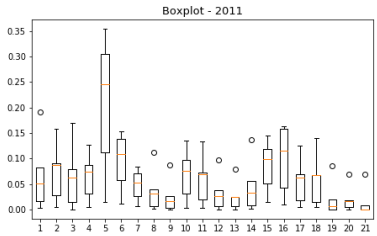
as novas features se relacionam com a classificação de mais alto nível, e caso seja muito diferente compara o agrupamento com essa novas varáveis aos feitos nesse trabalho.

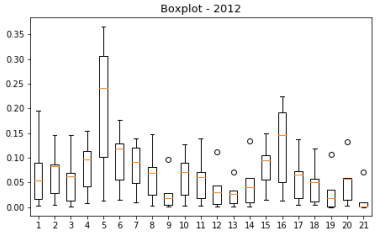
5 Anexo A - Boxplots

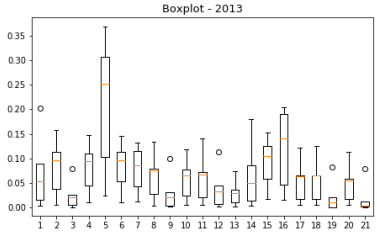


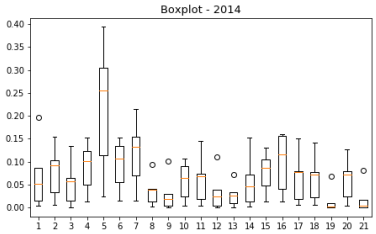




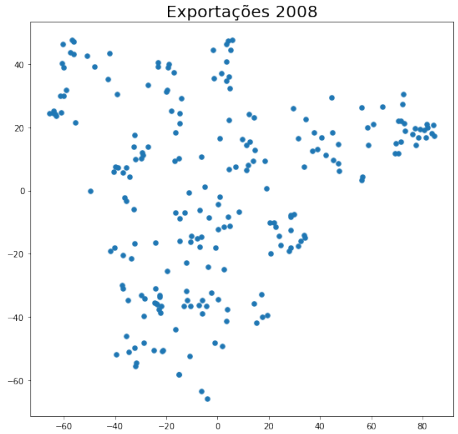


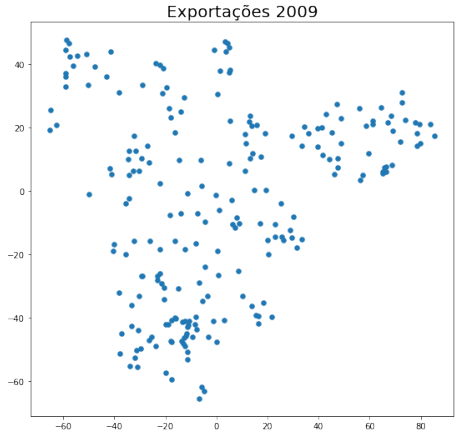


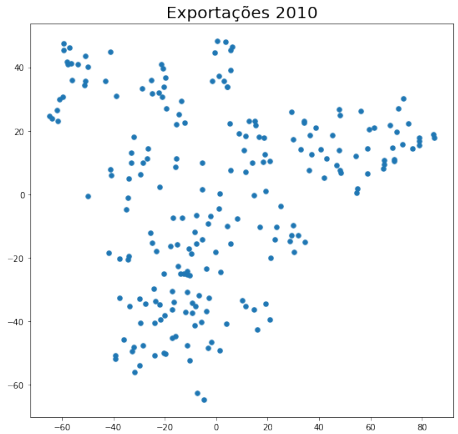


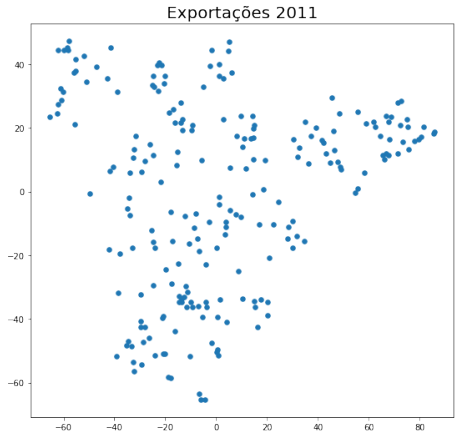


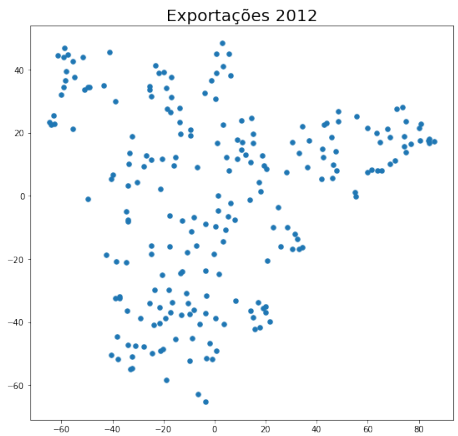
6 Anexo B - t-SNE

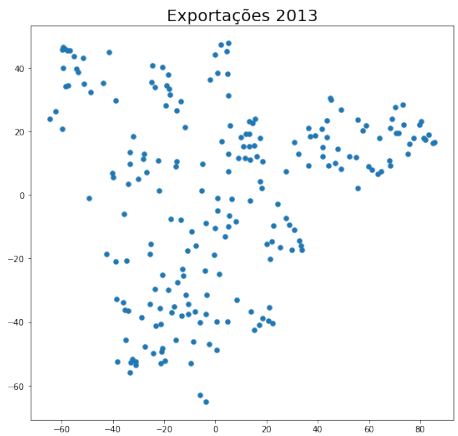

Exportações 2008

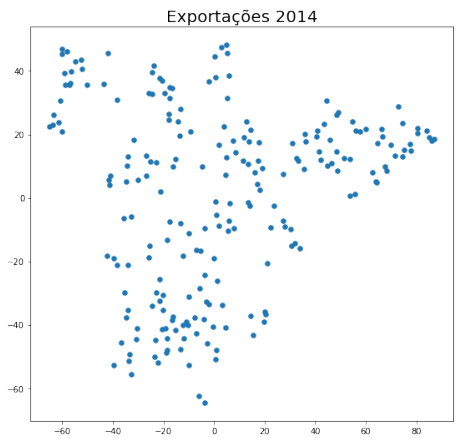

Exportações 2009


Exportações 2010


Exportações 2011


Exportações 2012


Exportações 2013


Exportações 2014