# Agrifood Data Study Group

## Evaluating the UK's Resilience to Supply Chain Shocks


Problem prepared by:

Alexandra Brintrup (University of Cambridge

Güven Demirel (Queen Mary University of London)

**Report Authors and Contributors:**

Martine Barons (University of Warwick), Alexandra Brintrup (University of Cambridge), Chris Budd OBE (University of Bath), Alan Champneys (University of Bristol), Po-Yu Chen (Imperial College London), Jessica Chen-Burger (Heriot-Watt University), Güven Demirel (Queen Mary University of London), Andrew Duncan (Imperial College London), Ali Fahmi (Queen Mary University of London), Lilia Georgieva (Heriot Watt University), Xue Gong (Maxwell Institute Graduate School), Samuel Johnson (University of Birmingham), Edward Elson Kosasih (University of Cambridge), Stephen Mak (University of Cambridge), Julie McCann (Imperial College London), Bart McCarthy (University of Nottingham), Stephan Onggo (University of Southampton), Yaniv Proselkov (University of Cambridge), Andra Sonea (University of Warwick), Cong Zhao (Imperial College London).

**Our thanks also to useful contributions from:**

David Brown (FareShare), Paul Clarke CBE (Ocado), Duncan Harding (Defra), Graeme Heron (University of Sheffield),

**This Virtual Study Group was hosted by the:**

Virtual Forum for Knowledge Exchange in the Mathematical Sciences (V-KEMS) consisting of Knowledge Transfer Network (KTN), Isaac Newton Institute (INI), Newton Gateway to Mathematics, International Centre for Mathematical Sciences (ICMS)




# Contents

# 1 Executive Summary

This food system underpins many aspects of our society. It feeds us and shapes the economic, social and natural environments that we live in. The COVID19 pandemic has further highlighted the importance of **resilience** in the food system [1].

This brief report examines the idea of the UK's resilience to shock by taking a data-centric approach. The following sections overview the progress of a group of mathematical scientists in evaluating the UK's resilience to supply chain shocks. Progress documented in this report is from a two day virtual Study Group held between the 7th and 8th September.

The aim of Section. 3 is mainly to illustrate how the methods of network science and statistics can be used to investigate the international dependencies of the UK's food supply chain and the resilience of the global food system. This analysis helps to identify the products and the countries that are the most critical among the UK's agricultural network dependencies and to understand how these evolve over time.

A suggested extension to the work would aim to exploit these network science methods to generate time series for individual country food stocks and flows, which will be used to study correlation networks that provides a tool for identifying critical flows

Section. 4 aims to evaluate the bottlenecks in our national food system. Two methods are identified and applied to ingredient level datasets to understand how product ingredients may be mined to understand dependencies both between ingredients and products. This approach might result in aggregate estimates of products that may be affected when the supply of a certain food ingredient is impacted. These approaches may also help us distinguish between essential and non-essential ingredients; and give us pointers to understand how food items can be replaced. Some possible further directions are discussion in relation to making a useful, practical predictive tool for players in the agri-food system.

Section. 5 begins to provide a mathematical language for describing how fair supply chains may be designed which provide utility for consumers, retailers, manufacturers and government. By focusing on a fair food supply to all, a toy model is provided which shows how ideas in mathematics and game theory might be useful for policy makers in understanding the role between fair food supply and agri-food policy.

# 2 Background

Food production, manufacturing, distribution and retailing is the UK's largest industrial sector at £121bn, accounting for 9.4 % of Gross Value Added (GVA) and over 4 million jobs. It has a huge environmental footprint accounting for 75 % of global freshwater consumption, 12MT of food waste, 24 % of anthropogenic carbon emission and 28 % of United Kingdom (UK) road freight by tonnage [1]. The complexity of the modern food system means that it is vulnerable to a range of shocks and stresses. Examples include a 2008 global food price spike (which saw the price of wheat rise by 130 %), a 2018 shortage of $CO_2$ gas that impacted supply chains across the food and drink sector, and panic buying in the 2020 COVID19 outbreak that led to shops introducing temporary rationing and limited opening hours.

In April 2020, the UK lockdown and developing pandemic disrupted amongst most other this, the normal activities of knowledge exchange. As such a group of national institutions and key mathematical scientists joined forces to provide a continuity service for mathematical science knowledge exchange and allowed for public, private and third sector organisations to access the discipline to pose questions around social distancing, track and trace, food poverty and many others.

An early topic for discussion was the food system and supply network and whether the mathematical sciences were able to provide some leadership and support in increasing robustness and resilience. An open forum with representatives of the food system was held on the 28 April 2020 [2] where three priority topics were decided:

1. How can the mathematical sciences support cooperation versus competition at different parts of the supply chain?
2. How to design optimal delivery systems which identify and prioritise vulnerable people?
3. How to design predictive models for upcoming bottlenecks, and using these to hit a dynamic target?

This report details progress made against the final priority.

To tackle the challenge, a two stage approach has been undertaken. In the first stage (this stage) preliminary data identification, proof of principles, and toy models were developed which go some way to assess the following:

**Identifying UK's dependencies to international food supply chains:**

1. What network representations can be made with available data on UK's global trade in the AgriFood sector? What levels of information can these representations provide us in

the context of UK's agricultural dependencies?

2. What products and geographical locations are the most critical among the UK's agricultural network dependencies to the rest of the world? How can these be reliably identified and quantified?

**Evaluating challenges and bottlenecks in our national AgriFood logistics networks:**

1. In the Covid-19 pandemic, which AgriFood products and UK regions have been most severely affected from disruptions to the supply and transportation of goods and labour?
2. What are the major bottlenecks in the national food logistics network and how can they be resolved?
3. How has the access of the low-income people to food been affected by the Covid-19 crisis?

Between the 7th to 8th September, 30 mathematical scientists (listed on page 3) met virtually to consider these questions. This report details their progress.

# 3 International Food System and the Dependencies of the UK Supply Chains

This part of the study focuses on the exploratory analysis of the international food system using secondary data on production, transportation, and trade. The global food system was studied in its entirety, which provided the context for understanding the UK's food dependencies on the international trade network in which it is embedded. In the first stage of the Data Study Group, the objective was to identify and pre-process relevant data in the public domain and illustrate the tools of the network science and statistics on this data.

Thanks to the availability of detailed production and trade data, an empirical analysis of the food system can be conducted at different scales, ranging from local to global networks and from granular individual product level to the level of more coarse-grained commodity groups.

Different questions can be answered by looking at networks at different levels, for instance the vulnerability to the disruptions in the supply of a specific product or the general food security. Therefore, different network representations were considered at different scales and several examples of agri-food product categories as well as local and global structures of the networks were analysed. A further consideration was the performance measures of interest, which can cover different dimensions of sustainability in addition to the dynamic measures of resilience and robustness that characterise the response of the food system to shocks.

Below, the aim is mainly to illustrate how the methods of network science and statistics can be used to investigate the international dependencies of the UK's food supply chains and the resilience of the global food system. This analysis helps to identify the products and the countries that are the most critical among the UK's agricultural network dependencies and to understand how these evolve over time.

## 3.1 Data

In this study, the FAOSTAT database [5] was used, the database of the Food and Agriculture Organization of the United Nations (FAO). FAOSTAT is a comprehensive international agri-food database, including national production of commodities, trade, prices, and also various indicators such as food security and environmental impact of agriculture. We used the **Detailed Trade Matrix** data set, which provides international agri-food product flows, and the **Food Balance** data set, which covers aggregate production, domestic consumption, import, and export levels.

These two data sets were supplemented by the **Product, Supply, and Distribution (PSD)** data set of the United States Department of Agriculture, Foreign Agricultural Service [6]. This data

set contains agri-food stock levels that is not provided in the FAOSTAT database; hence the two data sets were merged. Since the FAOSTAT and PSD data sets use different product labels and categories, these were mapped into larger commodity groups by using manual and string matching. After further validity checks with these mappings, which could not be completed due to the lack of time, this mapping will enable the generation of time series for individual country food stocks and flows, which will be used to study correlation networks that provides a tool for identifying critical flows.

A further aspect missing in the FAO data is the information on the environmental impact of food transport. For this, the data in Table. 1 and Table. ]reftable2 from the literature will be used in the further stages of the project.

Table 1: Modes of transport and emmission

| Mode | % of good freight in the UK | Avg $CO_2$-emission factor (g $CO_2$/TKM) |
|---|---|---|
| Roads | 90.5 | 62 |
| Railways | 9.4 | 22 |
| Internal Waters | 0.1 | 31 |
| Maritime (Short Sea) | N/A | 16 |
| Maritime (Deep Sea) | N/A | 8.4 |

Table 2: Estimated total distance

| Product type | Products of agriculture, hunting, and forestry; fish and other fishing products | Food products, beverages and tobacco |
|---|---|---|
| Roads | 12908 TKM | 38861 TKM |

We here visualise the geographical distribution of different commodities to demonstrate basic exploratory analysis with data. In Fig. 1, we plot the wheat stock variation in 2017 as a case of an important commodity. The map is colour coded with the per capita stock variation by country, which helps us to identify regions where the change in stock levels in 2017 was drastic. The countries in red and blue respectively indicate the regions where the wheat stocks decrease and increase. The shade of the color shows the magnitude of change and the countries with missing data are shown in white. Lithuania, Hungary, Canada, Kazakhstan, and Mongolia are the countries with most depleting wheat stocks in 2017, while Australia, Romania, France, Poland, and Turkey are the countries with highest increase in wheat stocks.

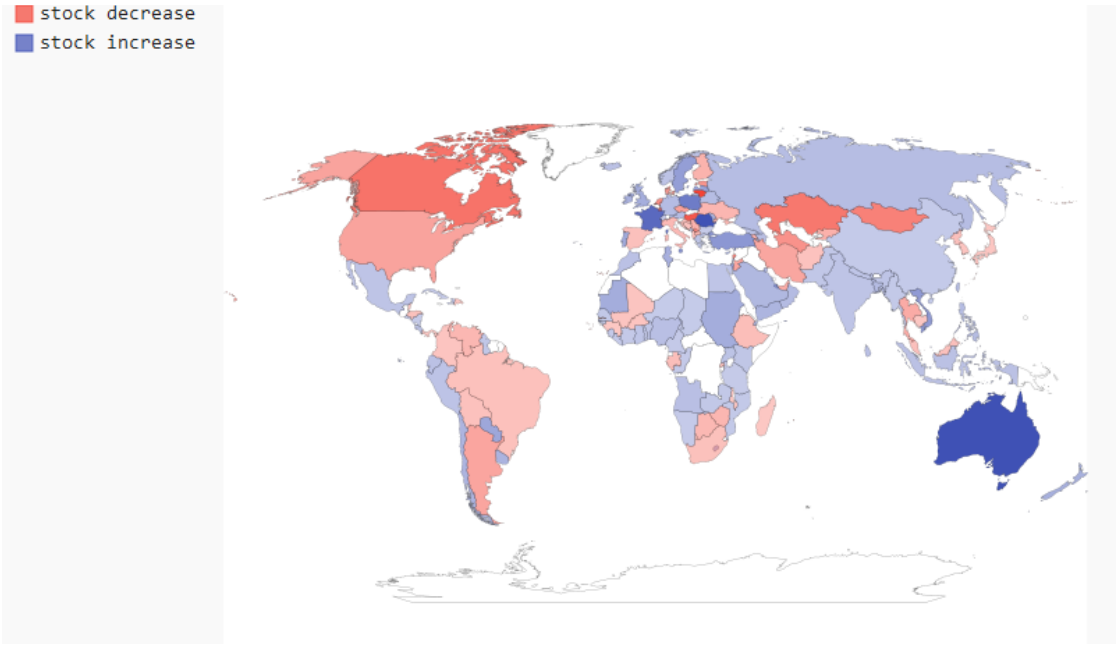


Figure 1: Wheat stock variation per capita (2017)

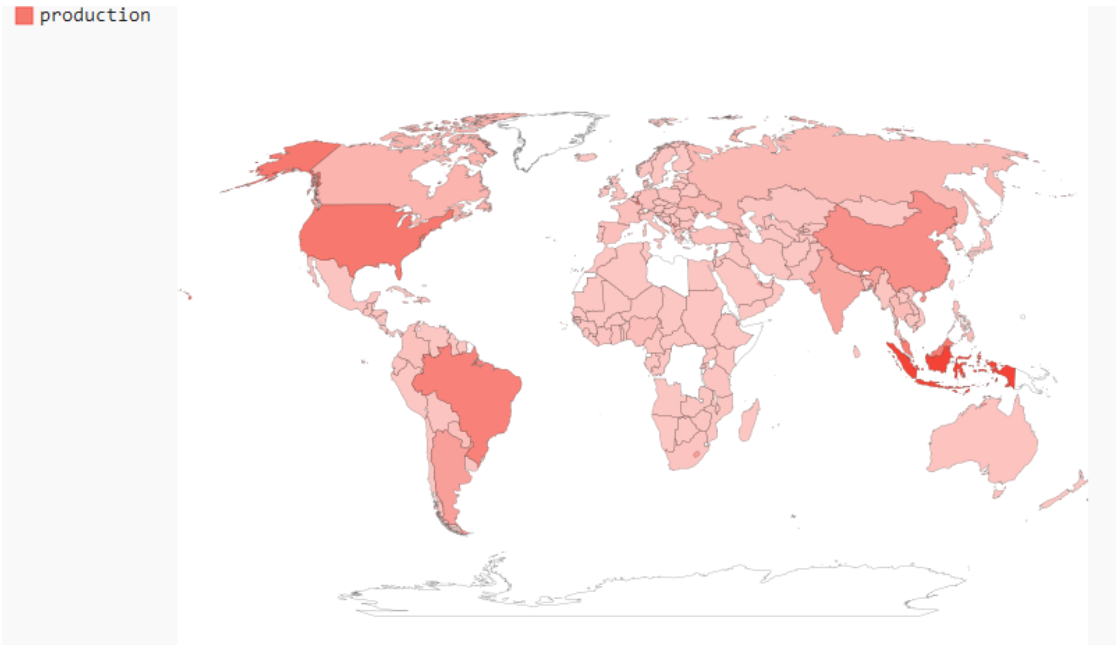


Figure 2: Oil bearing crop production per capita (2017)

Using the FAO Commodity Groups predefined mappings, and filtering out already present clustered data for each category, we can also apply the same geographical study on higher level categories, such as production in oil bearing crop and derived products. For instance, Fig. 2 shows the leading producers of oil-bearing crops in 2017.

We can inspect such graphs and detect changes in patterns over time. Together with basic statistical analysis, it provides summary information of the changes and generates some insights into vulnerabilities. We do not provide a detailed descriptive analysis here and instead focus on network analysis and visualisation of trade inter-dependencies. These tools complement such descriptive analysis and may help to discover vulnerabilities in the agri-food supply network that are not immediately evident in geographical distribution.

## 3.2 International food network analysis

The network analysis below involves several aspects; following data pre-processing, the first task undertaken was data exploration and visualisation. This was accompanied by an investigation into common measures of **network centrality** and **criticality** to understand the importance and vulnerability of individual countries within the global food trade network.

Furthermore, network properties, such as degree distribution and community structure, and their evolution through time can help us understand the global structure of the network. Different stakeholders may be interested in micro / meso / macro structure depending on their interests. The overall aim is to understand policies which help move towards an "optimal" network structure. However, resilience captures both economic, social, and environmental aspects; hence we need to consider different metrics simultaneously.

In the following, we consider calculating network-based proxies for diversity, redundancy, and resilience [10, 7, 8]. Such approaches enable us to calculate overall robustness / resilience of the system to shocks. Furthermore, they also enable us to identify which countries are particularly vulnerable. It is crucial to understand and predict future resilience levels; which can be assisted by using a combination of such approaches with statistical analysis of longitudinal data.

### 3.2.1 Cereal and cereal products trade network

In the below, we use FAO data on cereal to investigate the network structure of cereal trade. In the following, the edge weight represents the quantity of cereal and cereal products traded in 2018 between a pair of trading partners. The edges direct from the exporters to the importers.

### 3.2.2 Global cereal trade network

Fig. 3 shows the largest connected component of the global cereal trade network after retaining the routes with the highest trading quantities. The Louvain method was applied to the global trade network for community detection, and the colours of the nodes correspond to the corresponding communities. The network self-organizes into five main clusters by region; the UK is located in the Europe cluster (green). The size of the nodes represents the **weighted-out degree**, which corresponds to the total export quantity. The United States, Kenya, and Brazil are identified as the largest exporting countries.

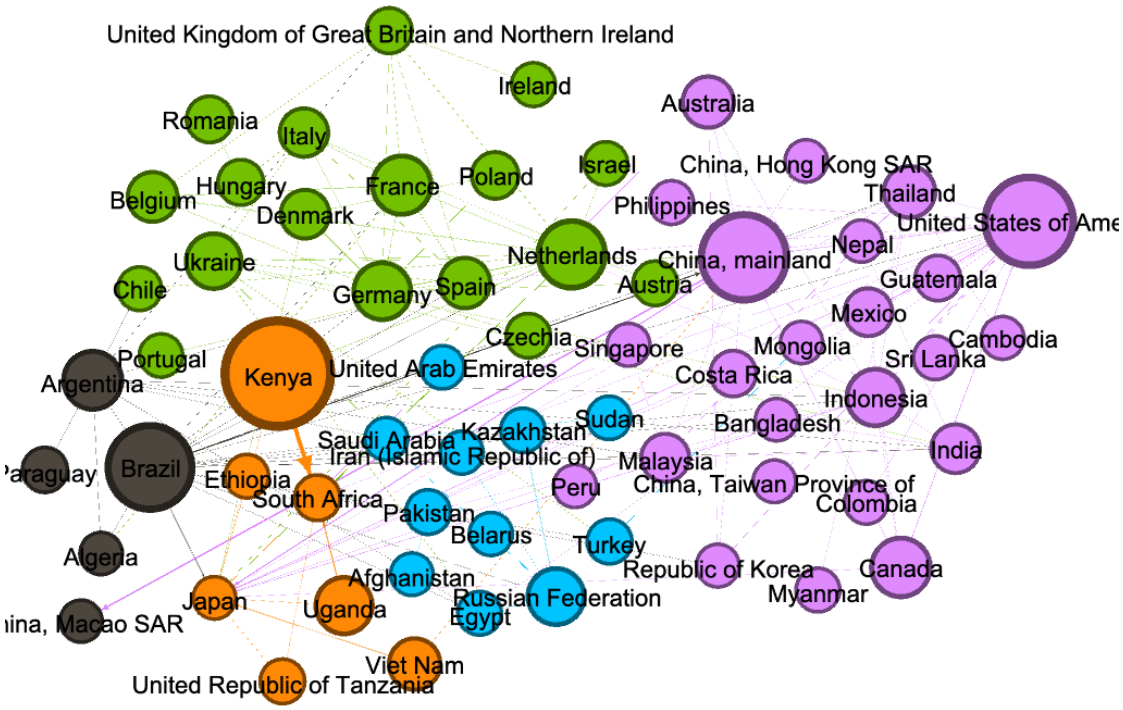


Figure 3: Global cereal trade network. Europe (green), Africa (orange), Asia-Pacific (purple), Middle East (blue), South America (black)

### 3.2.3 Regional cereal trade network

The European cluster is extracted from the global network in Fig. 3 for closer scrutiny. The European community can be further clustered into four groups using Louvain modularity, as shown by the colour of the nodes in Fig. 4. The size of the nodes represent the **weighted out-degree centrality**, or the trade quantity by weight. The Netherlands, Germany and France emerge as the top exporters in this region.

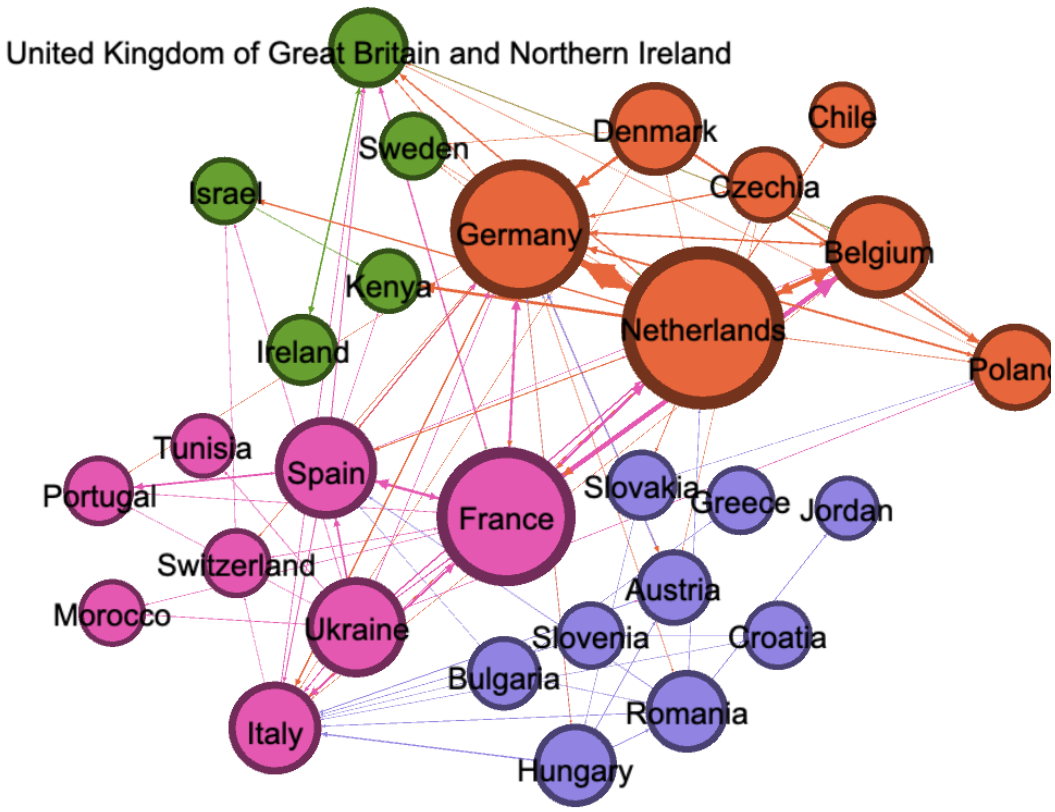


Figure 4: Regional EU cereal trade network - weighted out-degree centrality

The size of the nodes in Fig. 5 represent the **betweenness centrality**, this measures how many shortest paths pass through a node. Betweenness centrality can be interpreted as the importance of a country as a pathway in the trade network. If a node with high betweenness centrality breaks down, the trading routes could be severely affected, resulting in fragmented network or longer distances of the routes.

France, the Netherlands and Switzerland have the highest betweeness centrality. This is not surprising considering the central location of Switzerland and France in Europe, and the fact that Netherlands owns the largest port in Europe.

The above figures begin to demonstrate how network science representations can be used to describe global and regional structure of the food system, and use network metrics which can be interpreted in terms of supply chain robustness. Ultimately, we would wish to use these descriptions to inform policies which promote an "optimal" network structure. Of course, in reality, the optimal structure will need to be resilient to various constraints; economic, social and environmental.

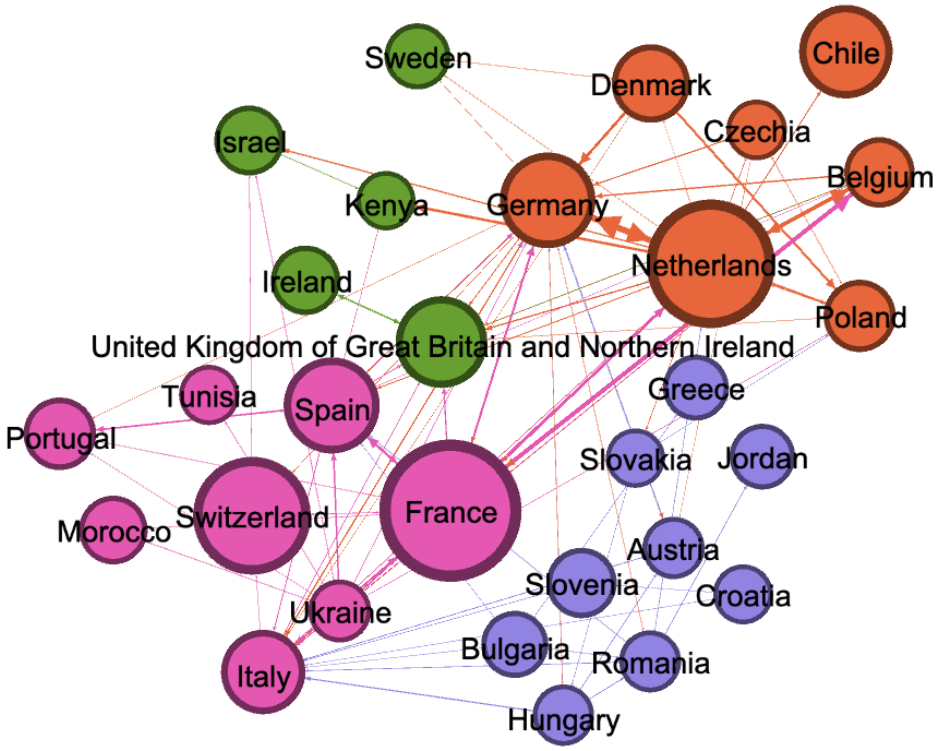


Figure 5: Regional (EU) cereal trade network - betweenness centrality

### 3.2.4 Detailed Trade Matrix

In this subsection, we have aggregated the data from the **Detailed Trade Matrix**, which tracks international trade flows for different food-related items. This dataset provides for each country the quantity and value (USD) of imports, export to all other possible countries of a standard set of items. The big challenge here is due to the large amount of missing data, which is due to the fact that the reporting country doesn't report all the possible trades every year, especially in situations where no trade is made. This raises the question whether or not a missing entry is simply zero or missing. We introduced a scheme for imputing missing data based on some intuition and statistics derived from the population of all edges. Based on this dataset, we can assess all the trade per year in and out of the UK.

To assess the diversity of sources of different produce, we compute, for every commodity, the ratio of trade from each country to the total imports into the UK for that commodity. This is equal to one if there is a single country exporting that commodity to the UK, and zero if it is only providing a small fraction of that commodity to the UK. Generally, this will be a number in between one and zero.

For a given commodity we compute the maximum ratio, this gives a measure of the **fragility**

with a ratio of one meaning it depends on a single country, and a ratio of zero meaning there are multiple countries supplying the commodity.

More formally, given countries indexed by $i = 1, \ldots, M$ and commodities indexed by $c = 1, \ldots, C$ define $F_i^c(t)$ to be quantity of commodity imported to the UK during year $t$. Let

$$p_i^c(t) = \frac{F_i^c(t)}{\sum_j F_j^c(t)},$$

the proportional contribution of the $i$th country to the UK's supply of commodity $c$. We then define the diversity score to be

$$D^c(t) = \max_{i=1,\ldots,M} p_i^C(t).$$

Clearly, $D^c(t) \in [0, 1]$ with 1 being obtained when the UK has a single supplier for commodity $c$ during the year $t$. As a demonstration we consider four items and assess imports into the UK, Fig. 6.

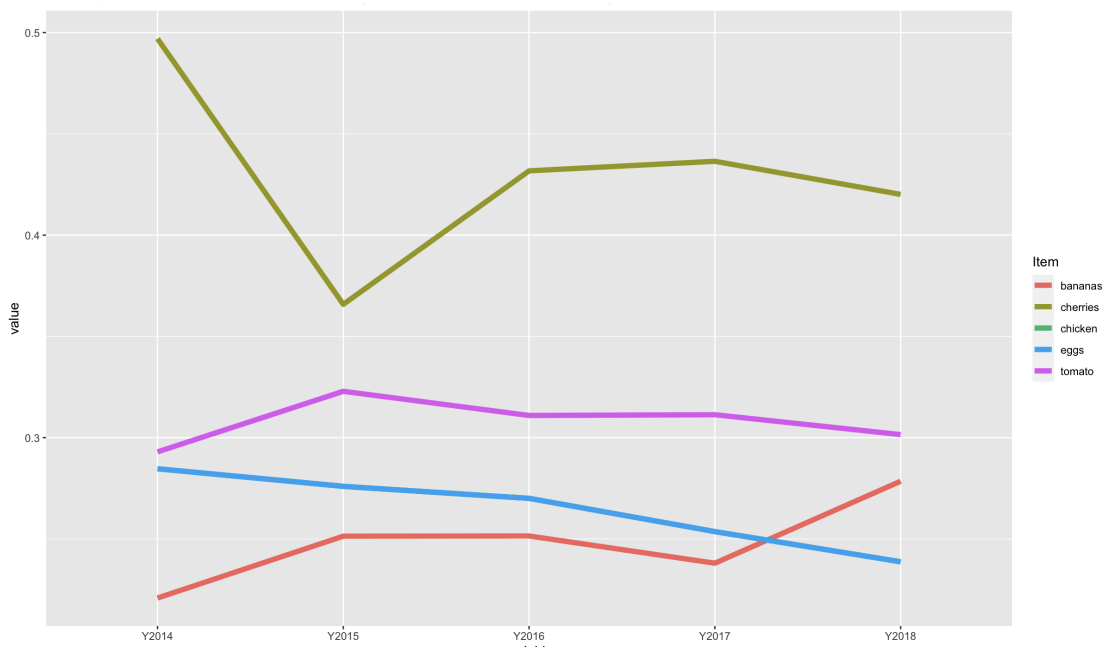


**Figure 6: Diversity score over time for items - the highest more dependent on a single trade flow.**

The hypothesis is that there is a correlation between the price per unit weight and the the diversity score. This can easily be tested using the detailed trade database. Of course to have a more complete picture, we need to integrate details of production, which can be done with Food Balance Data.

A more natural measure of diversity is the Shannon diversity index, which has been a popular diversity index in the ecological literature and is given by

$$H' = -\log \prod_{i=1}^{M} p_i^c(t)^{p_i^c(t)}.$$

The Shannon diversity index for four products is shown in Fig. 7). Why then is the diversity score of tomatoes (and derivative products) so low?.

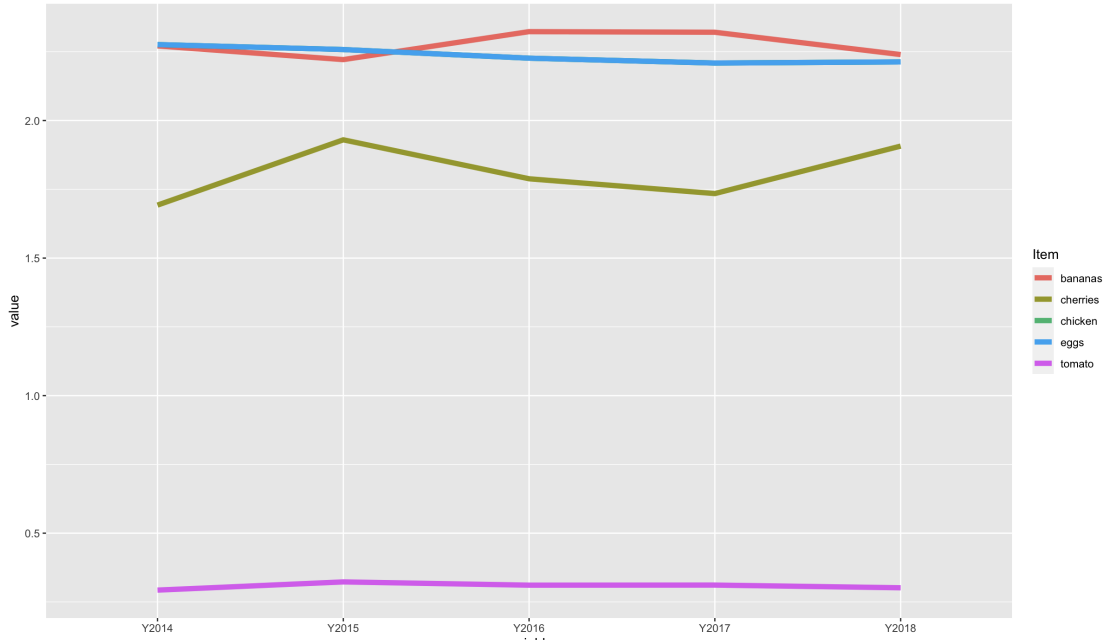


Figure 7: Shannon diversity score over time for items.

To answer this, we see in Fig. 8 that there are three strong spikes, namely Italy, the Netherlands and Spain which cover the total proportion of imports of tomato related products. On the other hand, Fig. 9 shows the same for bananas.

This suggests that there's a less strong domination on a single sets of countries over the years with regular imports from Columbia, Costa Rica, Dominican Republic, Equador and Belize, Cameroon, Panama and Ireland, etc.

We can conjecture that food supply diversity increases at the cost of increased trade dependence. This conjecture can be explored once we link to food balance data.

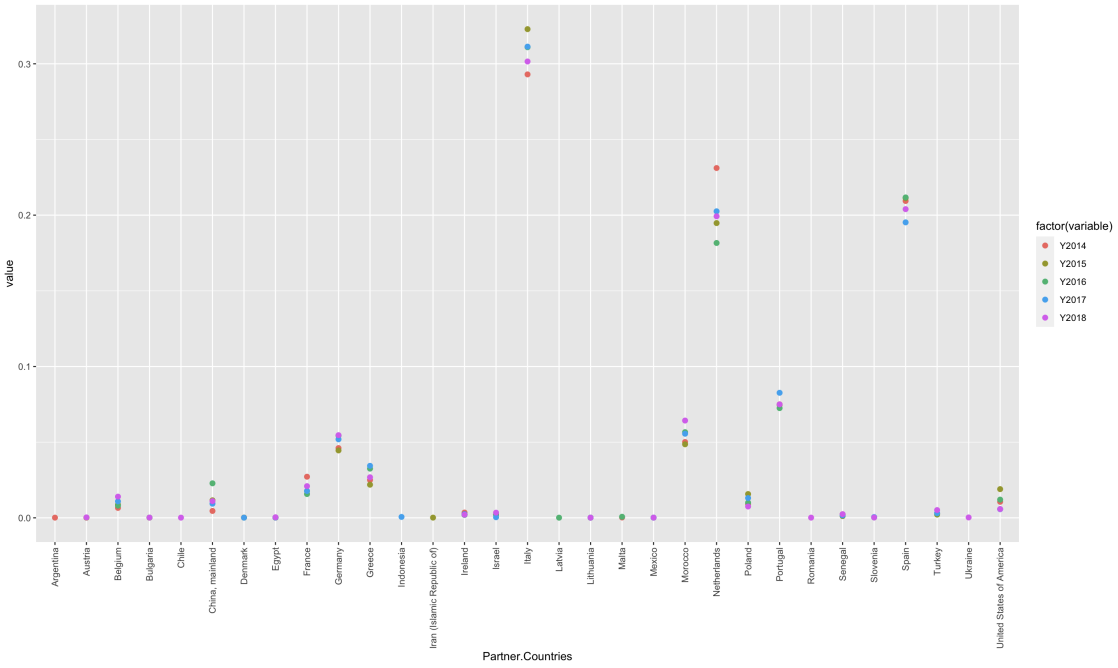


Figure 8: Imports of tomatoes by country

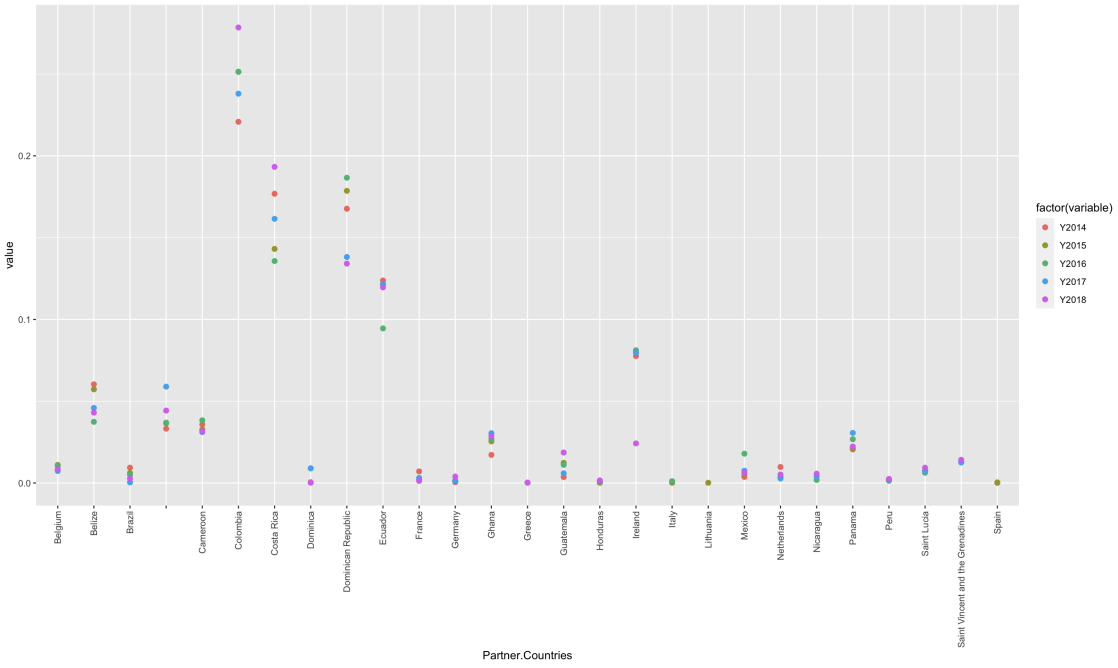


Figure 9: Imports of bananas by country

## 3.3 A dynamic Bayesian network model for food supply chain management

This final network representation uses Bayesian Network (BN)s. BNs are **directed acyclic graphs** consisting of variables and their dependencies. Each variable has a conditional probability table which parameterises the graph. A Dynamic Bayesian Network (DBN) is a temporal BN with a number of time slices. DBNs are Markov models which assume that the system is memoryless, that is the future state of a target variable depends only on its current state, not its previous states. We have collected a set of variables related to the food supply chain management in the UK. These variables include:

- Food Type,
- Packaging Capacity,
- Transportation,
- Food Import,
- Food Production,
- Food Export,
- Food Supply, and
- Emission.

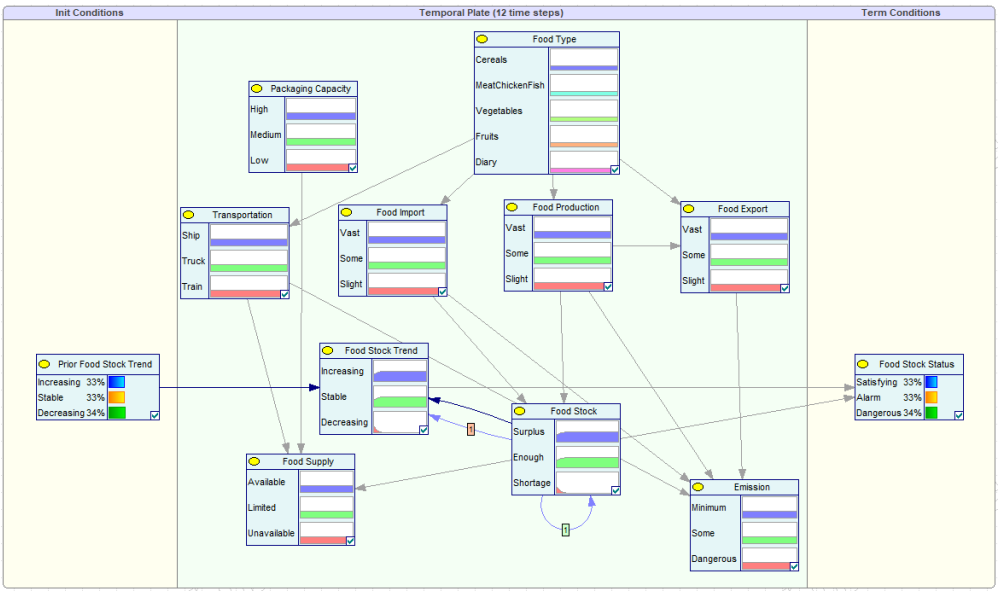


Figure 10: Final time slice of the DBN of a toy model of food supply chain management

Each variable has a set of states defined by the modeler. A trend variable is added to provide trend of food stock, indicating whether it is increasing, stable, or decreasing.

A toy DBN model for food supply chain management is shown in Fig. 10. It has a temporal plate which has 12 time slices of the model rolled in the plate. The target variable of this model is **Food Stock**. It has a first order connection which allows it to transit from one time slice to the next. The final outcome of the model is represented by the Food Stock Status variable, which combines Food Stock and Food Stock Trend variables of the latest time slice. Currently, the parameters are arbitrarily set.

An extension to this work would use real data to train the parameters of the model, after which the interpretation of the results will provide insights about the impact of different factors on the vulnerabilities of the food supply network and their evolution over time.

## 3.4 Summary and Next steps

This section has introduced a number of network and statistical tools to begin to assess potentially important metrics underlying a food systems resilience including diversity of supply, fragility of imports. Future activities should look to:

- use described network descriptions to define "optimal" network structure. Of course, in reality, the optimal structure will need to be defined in a way which includes not just economic, but social and environmental constraints.
- extend the Shannon diversity conjecture, to explore the relationship between food supply diversity increases and trade dependence. This conjecture can be explored once we link to food balance data.
- use real data to train the parameters of the DBN. This would provide insights about the impact of different factors on the vulnerabilities of the food supply network and their evolution over time.

# 4 Evaluating Challenges and Bottlenecks in our National Agri-food Logistics Network

This part of the study looks at how a data driven approach building on aggregated firm level data could help us prepare for shocks to the AgriFood sector. We started by briefly examining which agri-food products have been most severely affected from disruptions to the supply and transportation of goods and labour. The group discussed a number of cases, such as disruptions in the Pasta supply chain and demand vastly surpassing the rate with which products could be replenished on supermarket shelves.

Another case that was discussed was Unilever's Marmite - motivated by Media stories on how this typical British staple was only available as 250 g jars during the COVID-19 pandemic due to the scarcity of a key component: beer yeast sludge, a by-product of beer production, which went missing as large volume production of beer stopped due to pub closures. We went on to try to map supply chains in these cases with publicly available data as an exercise to understand what public data could tell us.

The data for these cases were scattered and not readily available. What was available were anecdotal evidence emerging out of recent shocks such as Brexit and the COVID-19 pandemic. While these give us some clues to the extent to which UK supply chains suffered disruption, a mathematical approach to preparing for potential shocks in the sector as a whole would need to be grounded in a data and scenario driven basis.

For example we know that around 55 % of our food is produced in the UK, but much of this food itself might be dependent on international food supply chains - yet we do not know how much dependency is there, in what food category, and how shocks to those dependencies would affect food production in the UK. Finding the right data is a challenging problem as firm level supply chain data is confidential and thus not accessible. On the other hand, available aggregated data from Office of National Statistics (ONS) such as sales in a particular product category is not fine-tuned enough to understand dependencies. Similarly, subscription databases that provide firm level data cannot be used. For example, in the Marmite case, one could only find supply chain dependency data on the corporate, (Unilever) level, which is again not fine-tuned enough as it would not tell us about the supply line of ingredients that make up Marmite.

The group then went on to examine how another data source can help. For this we looked at supermarket data on product labels with the aim of inferring potential dependencies. Both **network based analysis** and **machine learning** approaches were trialled. Our initial results show that dependency inference might be plausible, however more research is needed to improve the methods proposed, to predict dependencies at the food product level and to create

mechanisms to aggregate and link these to supply chain dependencies.

## 4.1 Dataset and approach

The goal of this exercise was to examine what dependency information can be extracted from a confidential ingredient dataset which was collected for a previous study by one of the authors from a private database. The dataset contained product categories and ingredients, gathered through the parsing of multiple instances of the same type of product to extract ingredients.

The data results in the extraction of a bipartite network with $I$ products and $J$ ingredients (Fig. 11). A weight $w$ can be attached to links to describe how many times an ingredient $j$ has been observed in the ingredient list of product $i$. The higher the $w$, the more confident we can be of the significance of the dependency between $i$ and $j$. However, we need to also be mindful that certain groups of ingredients such as salt might be ubiquitous, but not essential for the product to be produced.

Four questions arise:

1. How can meaningful and robust product-ingredient relations be extracted from such a dataset?
2. Can information arising from these relations be used to explore critical ingredients, the loss of which would yield most significant disruptions?
3. What measures can be devised to explore similarity between products so that they can be grouped for aggregate analysis?
4. Similarity could also be useful to predict which ingredients can be substituted to make the same product in case supply of a key ingredient becomes disrupted. What measures can be devised to explore similarity between ingredients?

To explore these questions, two approaches were followed. The first approach investigated **network based representation** (subsection. 4.2), and the second analysed the use of **natural language processing** clustering using two different methods (subsection. 4.3).

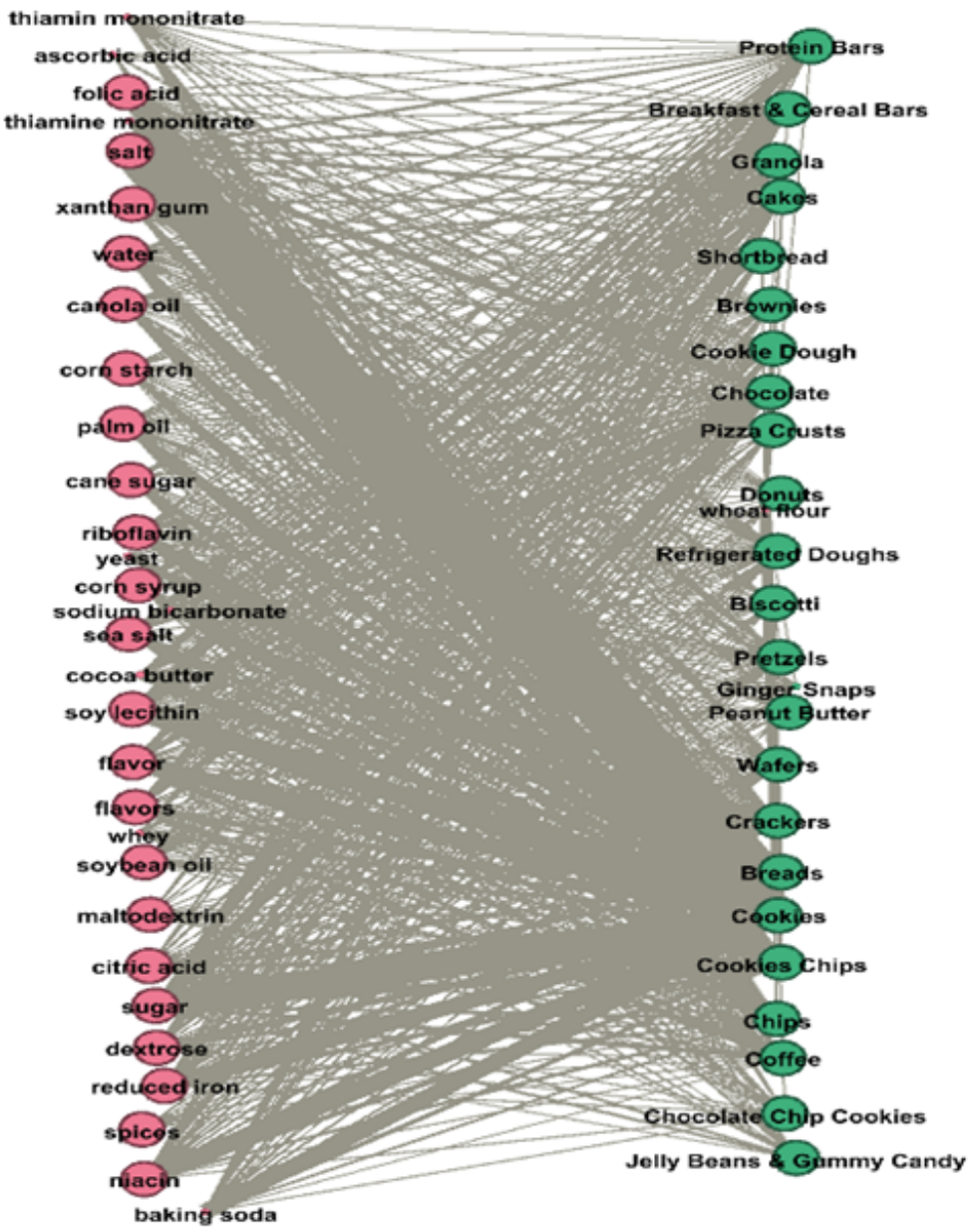


Figure 11: Product-ingredient network visualisation. Nodes on the left are ingredients, and nodes on the right are products. Products that share many ingredients are clustered together. Ingredient node size scaled to reflect the number of times it has been used in a product category.

## 4.2 Network science-based analysis

In the below subsection we look at how we might use **network science-based analysis** on the bi-partitie network introduced above. The aim of which is to understand from a data perspective ingredient importance and similarity.

### 4.2.1 Detecting similarity and importance

Let $A$ be an adjacency matrix describing a bi-partite network of ingredients and products (such as Fig. 11. That is, $a_{ij} = 1$ if ingredient $i$ is in product $j$. We could consider $A$ to be weighted. The number of shared products for two ingredients $i$ and $l$ is

$$n_{ij}^{P} = \sum_{l} a_{il} a_{jl} = (AA^{T})_{ij}.$$

The number of shared ingredients of two products $l$ and $k$ is

$$n_{lk}^{I} = \sum_{i} a_{ik} a_{il} = (A^{T}A)_{ij}$$

We can define two normalised adjacency matrices, $V$ and $W$:

$$v_{il} = \frac{a_{il}}{k_{i}^{out}},$$

and

$$w_{il} = \frac{a_{il}}{k_{l}^{in}},$$

where $k_{i}^{out} = \sum_{l} a_{il}$ and $k_{l}^{in} = \sum_{j} a_{jl}$.

The **out-degree** of an ingredient could be a measure of how important it is. The **in-degree** of a product would indicate its complexity. The symmetric matrices $VV^{T}$, $V^{T}V$,$WW^{T}$ and $W^{T}W$ provide measures of overlap between ingredients or products to which nodes contribute less the higher their in- or out-degrees.

### 4.2.2 Extracting product hierarchy

The network approach can be used to detect hierarchical pathways between ingredients and products. For example, cocoa is sold on its own as a product, however also appears as an ingredient in chocolate, which itself appears as an ingredient in a chocolate chip cookie, thus creating an inter-layer link at the ingredient level. **Hierarchical clustering algorithms** could be used to extract **dendrograms** that can determine appropriate clusters. Appropriate similarity measures and linkage criteria would need to be created to do so (Fig. 12).

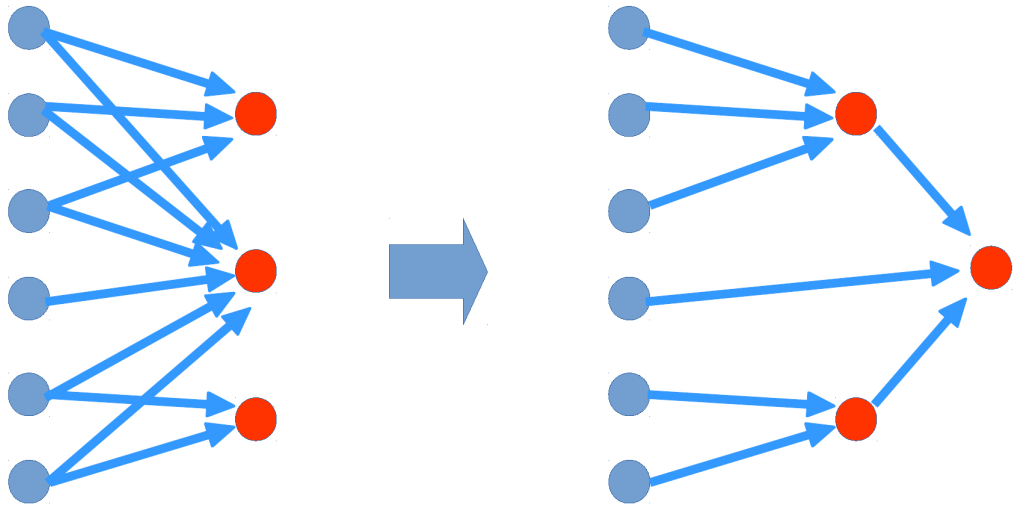

Figure 12: Hierarchical network product description.

Looking for paths in the network and distinguishing between essential and non-essential ingredients can tell us all the ways in which a product can be made.

### 4.2.3 Determining ingredient importance

As mentioned in subsection. 4.1 in the bipartite representation, the out-degree of an ingredient could be a measure of how important it is. The in-degree of a product would indicate its complexity. The symmetric matrices $VV^T$, $V^TV$, $WW^T$ and $W^TW$ provide measures of overlap between ingredients or products to which nodes contribute less the higher their in- or out-degrees. In a hierarchical network we would need to follow a slightly different approach:

- Here each node has a subset of nodes from which it receives input (its in-pyramid) and a subset to which it sends output (its out-pyramid).

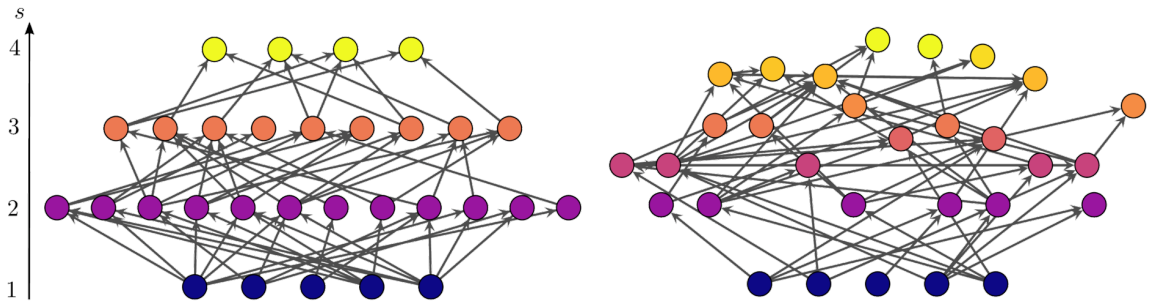


Figure 13: (left) Simple product hierarchy (right) more realistic, complex product hierarchy.

- Each node can be assigned a trophic level / product tier / Bill of Materials (BOM) level. If there are basal nodes (those with no in-edges) we can define the level of node $i$ as:

$$s_i = 1 + \frac{1}{k_i^{in}} \sum_j w_{ij} s_j,$$

  where $w_{ij} = 1$ means there is an edge from $j$ to $i$, and $k_i^{in} = \sum_j w_{ij}$ is $i$'s weighted in-degree. If there are no basals, we can use the method described in Ref. [17].

- We then identify "critical" components that are used in multiple end-products e.g. alcoholic drinks and hand sanitisers might depend on common components as these might be the bottlenecks in the network. Given limited resources for the latter, this might affect supply availability down the line.

Further, once similarity for every pair of products is identified (based on overlap of in-pyramids) one can deduce that the products which are common to similar nodes are more likely to be crucial.

## 4.3 NLP-based analysis

Here the aim of the analysis was to understand which products could be substitutable in the case of disruption. A **data clustering** experiment was performed on the dataset to understand whether product ingredient data could be used for such a scenario. The reason for using ingredients was that the list contained brand names, and ingredients were more generic, resulting in better match than the Google vocabulary which was used to train the clustering algorithm. Out of the 44,028 ingredients, 1,100 matches were found and the rest were discarded. The following approach was then used for analysis:

1. Each word is converted into a 300-dimensional embedding vector using the **word2vec**

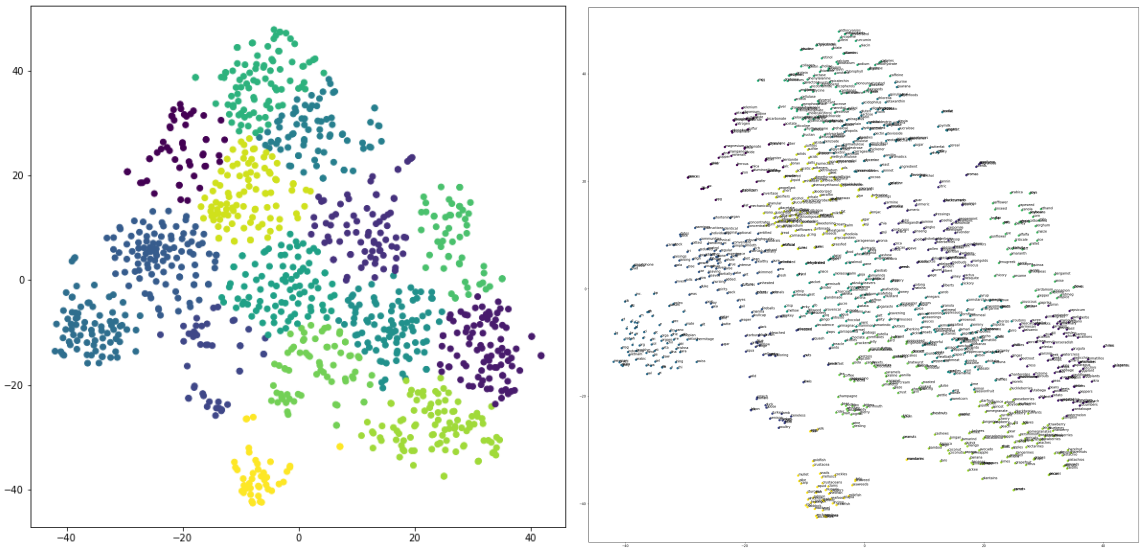


Figure 14: (Results of $K$-means clustering left) without label (right) with label

**neural network** which has been pre-trained on Google news containing approximately 3 million phrases.

2. A **Principal Component Analysis (PCA)** is applied to reduce the 300-dimensional vector into a 50-dimensional vector
3. **t-distributed Stochastic Neighbor Embedding (TSNE)** has been applied to reduce the 50-dimensional vector into a 2-dimensional vector
4. $K$**-means clustering** has been applied to cluster the 2-dimensional word vectors where the number of clusters has been specified as 15.

The resulting groupings are shown in Fig. 14. It is observed that the algorithm performs reasonably well where fish products, poultry products, alcoholic drinks, grains and seeds are successfully clustered. While this could be a proof-of-concept demonstration of the clustering approach, more fine grained ingredients have not been found in the training dataset and a large number of ingredient labels are missing - thus for a more informative cluster analysis models need to be trained on specific datasets pertaining to ingredient labels.

Furthermore, the **word2vec** model can be trained on an agrifood-specific text corpus, in order to capture semantic dependencies that pertain to our domain. This method is also called **transfer learning** in the field of machine learning.

## 4.4 Detecting similarity using feature extraction

Further analysis involved examining product and ingredient similarity separately and applying PCA. Given the dataset, product set are denoted by $\mathcal{P} = p_1, p_2, \cdots, p_l$. Given an ingredient $x$, we can create a $|\mathcal{P}|$-dimensional vector $v_x = w_1, w_2, \cdots, w_l$ where each entry $w_i$ represents the weight between this ingredient and the product in the dataset.

Each word / product is converted into an $N$-dimensional embedding vector (where $N$ is the number of unique ingredients in the database). Hence every dimension is an ingredient. By doing this, we can map each ingredient to a new space where each axis represent a product. In this space, the closer (e.g. Euclidean distance) the two points in this space, the more similar they are.

For example, if we suppose dim-1 is salt and dim-2 is flour, and that pizza and naan bread are close in dim-1 and dim-2 as they both have salt and flour, pizza and naan should be similar.

Since $v_x$ is usually very sparse, defining similarity in the above space can be inaccurate. Therefore, we apply two **unsupervised dimension reduction techniques**, namely PCA and TSNE to extract meaningful features by reducing the space dimension to a 2-dimensional space, where distances between ingredients can be representative.

This does not only provide more accurate similarity measure between ingredients, but also allows non-expert users to quickly visually identify the similarity between products or ingredients that may be interchangeable or highly correlated.

Fig. 15 (left) shows the preliminary results when applying PCA to reduce dimensionality to 50 and then applying TSNE to further reduce to a 2-D dimension. Fig. 15 (right) shows the preliminary results when only applying TSNE to the same dataset. It is almost comparable to the one obtained with PCA.

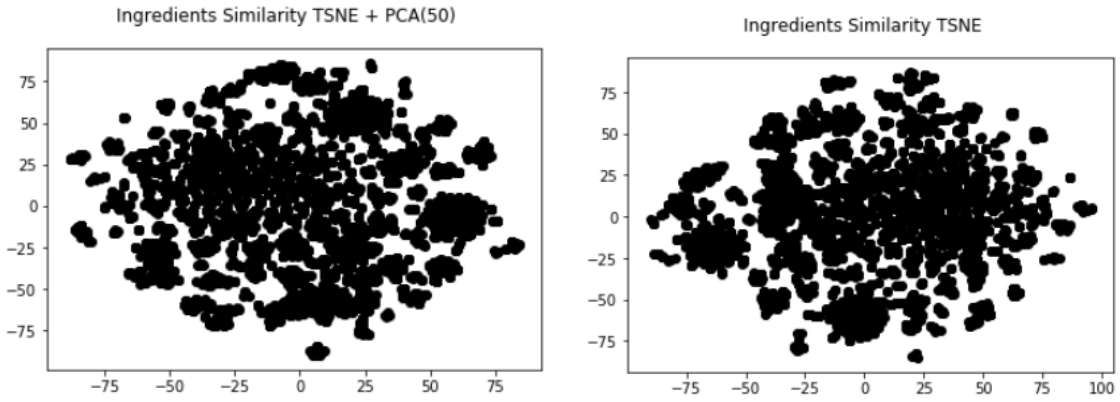


Figure 15: (left) without label (right) with label

## 4.5 Summary and Next steps

Our brief investigation into manually mapping food supply dependencies from public data pointed to the need for more robust, data driven methods to extract and analyse food supply. This is especially the case when complex food products made up of multiple ingredients are concerned. We decided to use an example, private food ingredient dataset supplied by Brintrup to analyse whether supermarket product labels can help in extracting useful information.

Two methods, namely directional bipartite network representation and clustering, have been identified to understand how product ingredients may be mined to understand dependencies both between ingredients and products; potentially resulting in aggregate estimates of products that may be affected when the supply of a certain food ingredient is impacted. These approaches may also help us distinguish between essential and non-essential ingredients; and give us pointers to understand how food items can be replaced.

Note that the algorithms applied were not fine-tuned and appropriate parameterisation was not carried out due to lack of time. Data that pertained the analysis was largely unprocessed and contained errors. Nevertheless, the very brief and preliminary analysis conducted here shows some promising directions which can be extended. Some further directions can include:

- Integrating multiple prediction approaches stemming from these approaches
- Incorporating further features into prediction
- Identifying food pathways through a hierarchical modelling approach
- Linking aggregate food import figures to product dependency information
- Modelling disruptions on the dependency chain

# 5 Designing "Fair" Agrifood Supply Chains

In the previous sections, we have used fairly simple measures of resilience, and not included environmental or societal considerations. The exercise outlined below takes a more holistic view of the concept. This section considers how utility can be gained by supporting the needy in our population and how to construct a system which shares utility around the various actors in the system. For example, can we understand how Government can use policy or public pressure to create a utility for the market to deliver "fair" supply to the needy? a schematic of interation and influence of the various players in the food system can be seen in Fig. 16.

The utility should be defined for each of the player in the food system. For Government, there are long term benefits in a fair food supply to all. These benefits might be lower healthcare costs relating to good public nutrition, public satisfaction with good food access could also be a political advantage. For retailers, utility can be in the form of positive social image due to a reduction in food waste and being more socially responsible. For many customers, a retailer with good social responsibility which support those needy in society, is a positive attribute. More widely, the benefit for the society and environment is the reduction in food waste.

To begin to turn this into a mathematical problem, we consider a single sub-sector or product such as pasta,ice or lettuce and develop utility functions for the various players; Government, retailers, customers, the needy, suppliers, distributors (such as FareShare). The functions are subject to constraints in money, total available supply of goods, nutrition etc.

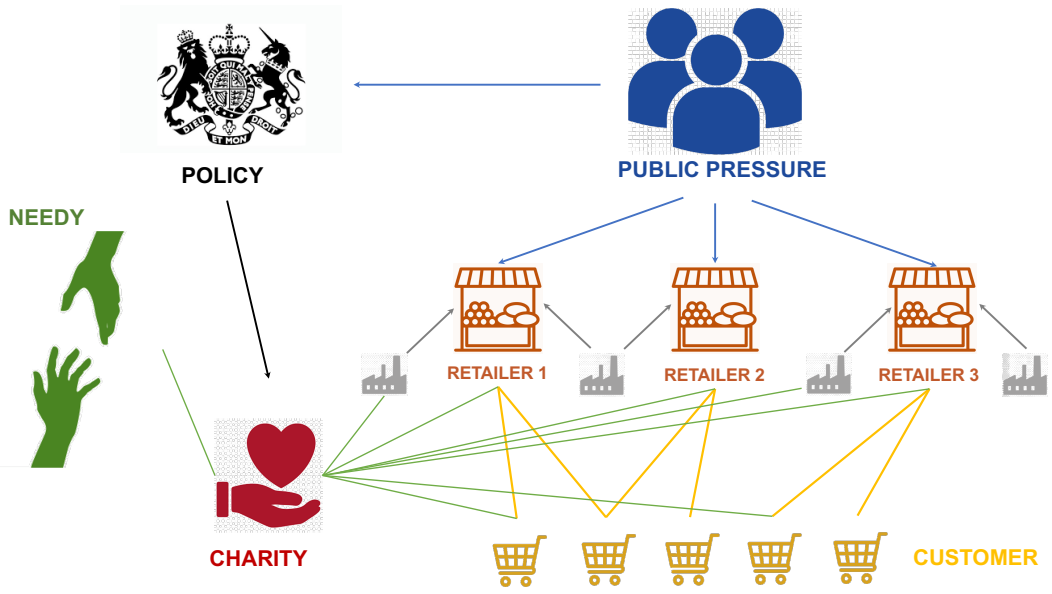


Figure 16: Interaction and influence between various players in the food system.

The variables which need be considered include; the demands from customers, the demands from the needy (who are unable to meet their own demands economically), profit of the retailer, profit of the supplier, cost to the consumer, cost to Government, cost to distributor, capacities (production/suppliers, distributors), and expected value of good (varies with time if unsold).

There are a number of factors which make such a problem complex; including time lag - both in terms of benefit (years) and in terms of time from intervention to outcome (weeks), forecasting both supply and demand, stochasticity including at a top level (e.g. weather), and spatio-temporal aspects.

## 5.1 A "toy" model for VCA 3.0

Let us consider a single product; for example lettuce, assuming that there is only one single supplier, and two kinds of consumer; a regular consumer $R$ and a "needy" consumer $N$. The needy consumer doesn't have access to the regular supermarket because of poor health or poverty, for example. We assume that there is an agent (e.g. FareShare) who seek to link the regular supermarket to the needy consumer.

We then want to construct a model with the various players in Fig. 16 which is a **multi-objective optimisation** which maximises the utility of the individual players collectively. This toy model considers some of the variables required to do this and provides a proof of principle demonstration as to the amount of lettuce $L$ going to the needy, being sold to regular customers, and being wasted and how much utility is gained in various scenarios.

Let us imagine discrete time $i4$, the weekly grocery provision and a demand from the regular and the needy consumer. In this toy model, the supplier wants to maximise price and the supermarket wants to maximise both its profit but also its reputation. We can build a utility function for the supermarket:

$$U_i = \sum_j [(P_{ij}(f_{ij}) - C_{ij})R_i] + G_i + \text{reputation}$$

where reputation could take the form of a function of product freshness, association with FareShare and amount of waste.

The simplest model of lumped supply. $L$ is known, and all the supply comes in on day 1. The amount of lettuce at time $i$ is then

$$L_i = L_{i-1} - \text{numbersold} - \text{numberwasted} - \text{numbersenttoFareShare}$$

| Variable | Description |
|---|---|
| $i$ | discrete time (in weeks) |
| $j$ | discrete crop (typically $O(100)$ lettuce) |
| $N, R$ | discrete number of groups of customers (typically $O(100)$) |
| $L_i$ | amount of lettuce in time $i$ |
| $f_{ij}$ | freshness of the $j$th lettuce at time $i$ on a scale of 1 (fresh) to 0 (inedible) |
| $N_i$ | needs of the needy |
| $n_i$ | proportion of $N_i$ actual supply to the needy |
| $P_{ij}(L_i, f_{ij}, R_{(i-k)})$ | Price at retailer of batch $j$. $P_i = f_0 + f_1$(demand-supply) |
| $C_{ij}(L_i)$ | cost (price at the supplier) |
| $R_i(P_{ij})$ | greedy desire of the regular customer. $R_i = \text{need} + \text{greed} - \text{const}\{\tanh \frac{\text{price}}{\text{const}}\})$ |
| $W_i$ | number of wasted at time $i$ |
| $W_i$ | $(R_i - R_0) + \max\{(L_i - R_i - n_i N_i), 0\}$ |
| $N_i$ | $R_0$ for some constant $R_0$ |
| $G_i$ | $f(W_i, N_i)$ Government and society intervention (costs of waste + incentive to feed the poor) $G_i = -\text{const} W_i + max(n_i - N_i)$ |

Table 3: Sensible parameters a toy optimsation model might include

What is the goal of the incentive? Supermarkets have the power. So, could we set this up as a competitive game to encourage the supermarket to compete. Goal of the government

- reduce waste
- enable sufficient supply to the needy
- minimize the (cost to the govt and to the economy of the) intervention

## 5.2 Model 0.1 implementation

For this simplified model of the above, the following further assumptions are made:

- there is no $j$
- $N_i$ is known (to the government but not the supermarket)
- Simplest case is 1 supermarket
- Government knows everything

- Design an intervention so that the supermarket maximise its utility (£) while also meeting the needs of the needy
- Decision variables are the fine for waste versus the benefit of meeting the needs of the needy
- Ignore reputation
- Even simpler fixed supply $L$ (case $L < R$(total), $L > R$(total))
- Timesteps; $i$ = 1, 2, or 3 for early, mid or late
- The freshness is a linear function of time $f_i = (4 - i)/4$
- For each time $i$ the supermarket needs to choose price $P_i$ and supply to needy $n_i$

We have implemented a version of this model. For simplicity we have assumed $price = freshness$ and cost $= 0.35$ per item. There are L items to sell in a week, any unsold items are waste. There is a penalty of 0.1 per item that goes to waste (in addition to the loss in being unsold).

In addition, there are $N$ needy people to be served by FareShare. There is a reputational and benefit (and/or government incentive of $d_1 N$ for using FareShare. But there is a penalty for not meeting the FareShare demands. So the utility to the supermarket is

$$U_{\text{FareShare}} = d_1 N - (d_2/N^2)[\sum_i n_i - N]^2 - d_3 n_i (1.0 - f_i)$$

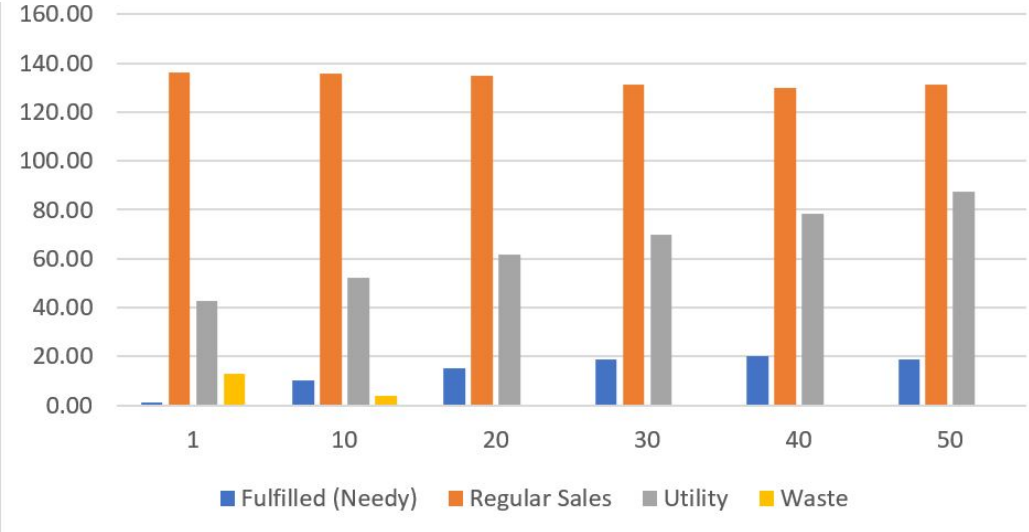


Figure 17: Model 0.1 outputs.

The values of $d_1$, $d_2$ and $d_3$ are the decision variables for government intervention. We have chosen $d_1 = d_3 = 1, d_2 = 10$.

$$U_{\text{total}} = L \times \text{cost} + \sum_i \text{profit} - 0.1 \times \text{waste} + U_{\text{FareShare}}$$

For an initial simulation we created $L$ = 150 a demand curve for the behaviour of regular customers that depends on freshness and price, and then run for different $N$. Using MS excel we have optimised the utility by allowing $n_i$, the amount going to the needy in each timestep to vary. The results are shown in Fig. 17

## 5.3 Summary and Next steps

Clearly, the above is a very simple interpretation of the food system, it does however show the kind of approach which could be taken to design **fair** supply chains. Extensions to the model would be to allow for a continuous supply model. Ultimately we would increase the number of supermarkets in the model and set up a competitive game between $N$ supermarkets and $m$ suppliers.

# 6 List of Acronyms

**BN** Bayesian Network
**BOM** Bill of Materials
**DBN** Dynamic Bayesian Network
**FAO** Food and Agriculture Organization of the United Nations
**GVA** Gross Value Added
**NLP** Natural Language Processing
**ONS** Office of National Statistics
**PCA** Principal Component Analysis
**PSD** Product, Supply, and Distribution
**TSNE** t-distributed Stochastic Neighbor Embedding
**VCA** Value Chain Analysis
**UK** United Kingdom

## References

[1] https://post.parliament.uk/research-briefings/post-pn-0626/

[2] https://ktn-uk.org/news/ktn-forum-on-agri-food-logistic-threats-and-research-opportunities-outcome-and-slides-now-available/

[3] https://code.google.com/archive/p/word2vec/

[4] Boccaletti et al. 2014 Physics Reports 544(1), 1-122

[5] Food and Agriculture Organization of the United Nations. 2020. FAOSTAT Statistical Database. [Rome]: FAO.

[6] United States Department of Agriculture, Foreign Agricultural Service. 2020. Market and Trade Data, PSD Online.

[7] Kharrazi A, Rovenskaya E, Fath BD (2017) Network structure impacts global commodity trade growth and resilience. PLoS ONE 12(2): e0171184

[8] Kummu et al, 2020 Global Food Security

[9] Seekell et al. 2017 Environ. Res. Lett. 12 025010

[10] Suweis et al 2015 PNAS 112:22 6902-6907

[11] Tu, C., Carr, J., Suweis, S. 2016. PLoS ONE 11(11), e0165941

[12] Wu,Z. at al A Topological Analysis of Trade Distance (2020) DOI: 10.3390/su12093511

[13] Thierry Mayer Soledad Zignago , 2011."Notes on CEPII's distances measures: The GeoDist database,"CEPII Working Paper 2011- 25 , December 2011 , CEPII.

[14] ECTA/CEFIC - Guidelines for Measuring and Managing CO2 Emission from Freight Transport Operations

[15] Feeding Vulnerable People. Study Group Report (2020)

[16] Agrifood Data Study Group: Evaluating the UK's Resilience to Supply Chain Shocks. 07/09/2020. Problem Statements & Contexts

[17] R. S. MacKay and S. Johnson and B. Sansom (2020), How directed is a directed network? https://arxiv.org/abs/2001.05173

Connecting for
Positive Change

**Head Office**

**KTN**
Suite 218, Business Design Centre
52 Upper Street
Islington
London N1 0QH

Telephone: 03333 403251
Email: enquiries@ktn-uk.org
ktn-uk.org
@KTNUK

**Contact Person**
**Dr Matt Butchers**
matt.butchers@ktn-uk.org