A Brief Analysis of Irish and European Crop Production



Group ID - MSc in Data Analytics

Author: Conor Dillon

e-mail: [sba22446@student.cct.ie](mailto:sba22446@student.cct.ie)

Student ID: sba22446

Date: 11/01/2023

**Contents**

[Abstract 3](#_Toc124371345)

[1. Introduction 4](#_Toc124371346)

[2. Project Management 4](#_Toc124371347)

[2.1. Stages of SEMMA (Quantum, 2019) 4](#_Toc124371348)

[2.2. Datasets 4](#_Toc124371349)

[2.3. Code 5](#_Toc124371350)

[3. Crop Production 6](#_Toc124371351)

[3.1. Historical Crop Production 6](#_Toc124371352)

[3.2. Geospatial Considerations 10](#_Toc124371353)

[3.2.1. Rainfall 11](#_Toc124371354)

[3.2.2. Policy 12](#_Toc124371355)

[3.3. Crop Yield Prediction 12](#_Toc124371356)

[3.4. Dashboard 13](#_Toc124371357)

[4. Crop Trade 14](#_Toc124371358)

[5. Sentiment Analysis 16](#_Toc124371359)

[5.1. Sentiment 16](#_Toc124371360)

[5.2. Test using Summarisation with GPT3 17](#_Toc124371361)

[6. Conclusion 18](#_Toc124371362)

[7. References 19](#_Toc124371363)

[8. Appendix 20](#_Toc124371364)

[8.1. GPT3 Summary of the Cereal chapter 20](#_Toc124371365)

Word count – 3541 (not including references and appendix)

# Abstract

This brief report introduces the topic of crop production for various countries in the European Union. Ireland’s top crops in terms of production are presented, with the top three selected for further analysis. Barley, wheat, and potatoes are shown to be the top produced crops. A comparison of these crops are then made between Ireland and a selection of other EU countries in terms of production, yield, and area harvested. Ireland’s is shown to on the lower side of production, primarily due to the much are used to grow and harvest. However, Ireland is shown to have the highest yield in all of these crops of any of the selected countries.

A choropleth was drawn for each of these crops for all EU countries. This revealed that there appears to be a geographical element to crop yield – countries of higher yield tended to neighbour the same. With this theory in mind, a semivariogram was plotted to numerically investigate this theory. Indeed, the semivariogram indicated that there is less variance in the yield data of countries that are closer together. This indicates a geospatial aspect to crop yield.

Precipitation is then considered as being the likely element of geography that is impacting crop yield. To test this theory, rainfall in each EU countries is plotted against crop yield for barley. A Pearson Correlation test was conducted to evaluate the relationship. The test returned a correlation of 0.55 with a p-value of 0.003. This represents a moderate relationship with a low probability of sample bias. Furthermore, a t-test was run to check mean yields for countries above and below EU median rainfall. The t-test returned a p-value of 0.006, indicating that the null hypothesis of the means being similar is rejected. Rainfall is indicated as a factor that affects crop yield.

To test a non-geographical factor for crop yield, government expenditure in agriculture is compared against crop yield for the years (2001-2019). Visually assessing the generated scatter plot revealed that there was no obvious correlation between barley yield and expenditure. An ANOVA test was also run to quantify the variability in the means. Specifically, a Welch ANOVA test was run as the variance in the data was deemed to be too high for a standard ANOVA test.

Ireland’s timeseries for wheat yield was found to have a high correlation (0.94). It is determined that a linear regression could be used to forecast future crop yields for wheat in Ireland. The R2 score returned is 0.89 and the model is deemed satisfactory. Barley yield is predicted to be 103,494hg/ha by 2050, up from 78,608hg/ha in 2019.

A dashboard is created to present the choropleth data for barley, wheat, and potatoes.

The Irish trade crop trade data is also investigated. A theory is put forth that crop production has an effect on crop trade price. A pairplot reveals that an increase in production leads to an increase in export and a decreased in import quantities. A price/production scatter plot is constructed and a K-means clustering model is employed to identify which country points belonging to each country.

Finally, a sentiment analysis is performed on an agricultural review annual produced by the Department of Agriculture, Food, and the Marine. The chapter on cereals is analysed using a natural language processor and a breakdown of the sentiment score is plotted. As a means of reviewing the result, OpenAI’s API is utilised to provide summarisation of the same chapter. The analysis concludes with a positive outlook on the industry with the natural language processor agreeing with the sentiment as assessed by the Author.

# Introduction

This report explores crop production in Ireland and the EU. As a subscriber to the EU’s Common Agricultural Policy, Ireland benefits from a range of advantages including financial support, food security, rural development, trade, among others. CAP does present some disadvantages to member states. Foremost, CAP is an expensive policy, costing EU taxpayers €58 billion per year (Zahrnt, 2011). Investment and supply in fertilisers and pesticides have led to a negative impact on the environment. CAP has also been criticised for essential effecting trade barriers between the EU and non-EU countries.

A successful enforcement of these policies requires good data. For each country in the EU, their department of agriculture is required to gather data across a range of topics and report back to the European Commission for the effective monitoring of Industry. A side-effect of this process is the general public’s access to high quality data covering a myriad of topics. This is the data that is used for this analysis.

# Project Management

The project management framework chosen for this project is the SEMMA (Sample, Explore, Modify, Model, and Assess) strategy. Distinctly cyclical in nature, this framework is chosen as it is expected that the project could become iterative. It was presumed that initial data exploration and analysis would propose theories leading to further exploration and analysis, and so on.

## Stages of SEMMA (Quantum, 2019)

**Sample** — a portion of a large data set is taken that is big enough to extract significant information and small enough to manipulate quickly.

**Explore** — data exploration can help in gaining understanding and ideas as well as refining the discovery process by searching for trends and anomalies.

**Modify** — data modification stage focuses on creating, selecting and transformation of variables to focus model selection process. This stage may also look for outliers and reducing the number of variables.

**Model** — there are different modelling techniques present and each type of model has its strengths and is appropriate for a specific goal for data mining.

**Access** — this final stage focuses on the evaluation of the reliability and usefulness of findings and estimates the performance.

Examples of each stage in the context of the proposed questions are presented in the following sections.

## Datasets

Data in this report was sourced from the Food and Agriculture Organization of the United Nations (FAO). Their FAOSTAT tool allows the user to select a range of data types and using several types of filters e.g. countries, items, years, etc.

cropProduction.csv contains information on yield, production, and area harvested for various crops for EU countries.

cropTrade.csv contains information on quantity and value for various crops traded by EU countries.

average-precipitation-per-year.csv contains the average yearly rainfall for many countries. This was sourced from worldbank.org.

## Code

The code for this project was written in four notebooks

* AgirExpenditure.ipynb (left misspelled to preserve commit history)
* cropProduction-Copy1.ipynb
* Sentiment Analysis.ipynb
* Dashboard.ipynb

The repository is public and is available at <https://github.com/ConorD-CCT/CA2_repos>.

# Crop Production

## Historical Crop Production

For the comparison of crop production between Ireland and other EU countries, it was decided to reduce the number of crops to three. Analysis of the FAO data shows that Barley, Wheat, and Potatoes are Ireland’s top three produced crops in terms of tonnage, Figure 1.

Chart, bar chart, histogram

Description automatically generated

Figure Ireland's top ten produced crops (in tonnes)

In Figure 2, Figure 4, and Figure 6, there is a comparison between Ireland and number of other EU countries in terms of the production, yield, harvested area for barley, wheat, and potatoes. A selection of the EU’s largest economies are chosen for comparison with Ireland – Netherlands, France, Italy, Spain, Poland, Portugal, Germany, Sweden, and Finland. Production refers to the amount of crop produced in tonnes. Area harvested refers to the area of land that the crop is produced from. Yield refers to the amount of crop extracted per unit area. Thus, a crop’s yield could be considered a better measure of a country’s success or efficiency in producing the crop. A common theme for most countries is the gradual increase in yield since 1961. This is likely due to improvements in farming technology, methods, and the increase in use of high-quality fertiliser. For each of the crops, Ireland is recorded as having the highest yield value in 2021. Section 3.2.1 examines these statistics in order to determine if Ireland’s climate is contributing to this success.

**Chart

Description automatically generated**

Figure Barley

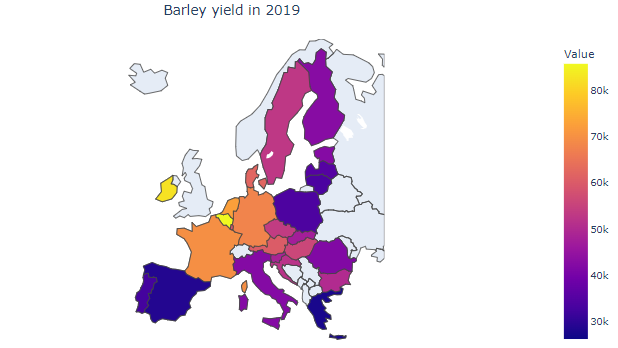


Figure Barley choropleth

Chart, histogram

Description automatically generated

Figure Wheat

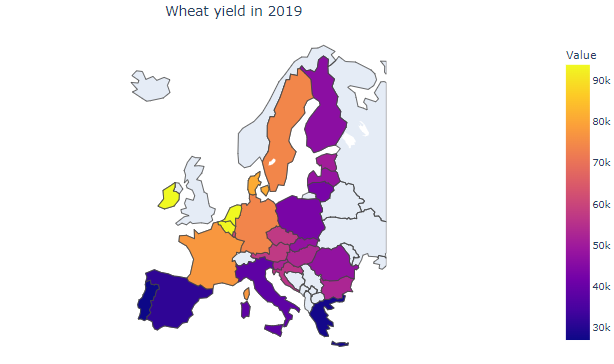


Figure Wheat choropleth

A picture containing graphical user interface

Description automatically generated

Figure Potatoes

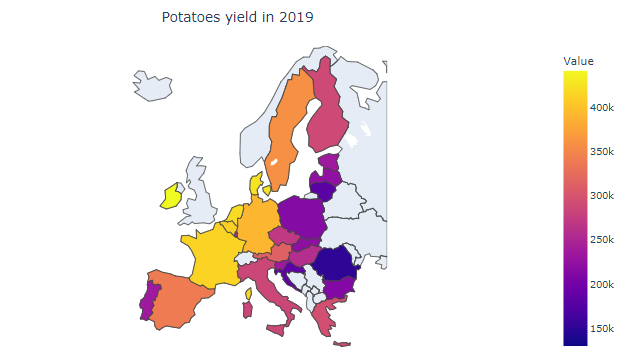


Figure Potatoes choropleth

## Geospatial Considerations

Chart, line chart

Description automatically generated

Figure Semivariogram for Barley Yields for Europe in 2019

A semivariogram has been plotted to model to effect of geospatial effect on crop yield in Europe, Figure 8. The theory is that the property that is causing higher yield in some countries is due to a geographical effect as opposed to a country’s agricultural policy. Examples of this could be precipitation, sunlight, soil quality, etc.

The function findGeoCode() was written to acquire the latitude and longitude of the each country and was added to the dataframe. This dataframe is then passed to makeSemivariogram() to make the semivariogram.

The plot shows a trend downwards at lower lags, indicating that lower distances can lead to lower semivariances. There are some outliers in the calculation at higher lags. The model would benefit from the inclusion of a larger amount of countries.

The next two subsections test further if the occurrence of higher crop yields is due to geographical reasons and/or policy. The first subsection tests if higher crop yield may be as a result of higher precipitation. The second investigates if higher government spending in agriculture leads to higher yields. The dataset is reduced to consider barley only – Ireland’s highest produced crop.

### Rainfall

From the previous section, it is suspected that there are geospatial considerations when trying to understand crop yield between countries. It theorised that a country’s rainfall has an impact on crop yield. To test this, rainfall is plotted against yield.

Average yearly precipitation data was sourced from WorldBank.org. The figure taken is the average yearly rainfall between 2010 and 2019. Taking all EU countries, yield of barley was plotted against average yearly precipitation in a bid to determine if there is a correlation between barley yield and rainfall.

Chart, scatter chart

Description automatically generated

Figure Barley yield vs rainfall in 2019

Figure 9 shows a moderate relationship between yields and rainfall for each country. Running a Pearson Correlation function returns a correlation value of 0.55 and a P value of 0.003, meaning that the probability of the result being a result of this specific sample being correlated due to chance is 0.3%. Variability in the data may be the result of many factors such as soil quality, amount of sunlight, fertiliser use, seasonality of rainfall, etc. That said, it is reasonably determined that an increase in rainfall would lead to an increase in yield.

A hypothesis is put forth that higher than average rainfall leads to higher crop yields. To test this hypothesis, a dependant samples t-test is used. The categorical descriptor is a variable indicating if that country’s rainfall is higher than the EU median. The numerical value is crop yield. The t-test is performed using the scipy stats module and returns a P value of 0.006. The t-test results with the determination that the difference in crop yields between the lower-than-median and higher-than-median rainfall countries is statistically significant – the chance that the difference due to the sample selection of the sufficiently low. The test infers, for a wider population of countries, that higher rainfall leads to higher crop yields. The null hypothesis is rejected. Historical rainfall would benefit this analysis, but it was not available.

### Policy

In order to test if government policy is a contributing factor to crop yields, it is decided to investigate government expenditure in agriculture. An expenditure dataset is acquired from FAO and merged with a dataset containing information on EU GDP. To normalise the data, government expenditure is taken as a percentage of GDP and compared against crop yields.

An ANOVA test is undertaken to measure if there is a statistical difference between the yields of barley for each country. A Shapiro-Wilk test is performed on each of the samples to check for normality. It is found during these tests that the Netherlands sample is not normal. The Netherlands sample is excluded from the ANOVA test. A Levene test is undertaken to check the variance of the samples. The Levene test returned a p-value of 1e-5, meaning that the variances are not equal. As a result, a Welch ANOVA test must be used. The Welch ANOVA test returns a p-value of 1.1e-3, indicating that the sample’s means are statistically different. The null hypothesis is rejected.

Chart, scatter chart

Description automatically generated

Figure Barley Yield vs Agri Expenditure (2001-2019)

## Crop Yield Prediction

There appears to be a strong upward trend in the Ireland’s historical barley yield, Figure 2. The correlation of this trend is determined as 0.94. Given the linearity of this data, it is considered feasible to perform a linear regression on this data to forecast future barley yields for Ireland. The module used is the linear regression model from the sciket-learn library. The data is scaled using the *MinMaxScaler* and the data is split into 70% train and 30% test using the *train\_test\_split* function.

After the data is fit to the train data, a prediction is made for the train data. The R2 score returned is 0.89 and the model is deemed satisfactory. Barley yield is predicted to be 103,494hg/ha by 2050, up from 78,608hg/ha in 2019.

Chart, scatter chart

Description automatically generated

Figure Barley Yield Linear Regression - Test vs Prediction

After determining the coefficient and intercept, the equation of the line can be written as:

## Dashboard

A dashboard was constructed to show the yield of several crops on a choropleth of Europe. The dashboard is constructed using the Dash library. The dashboard uses a drop down menu to select between the three crops that have been analysed in the previous sections. There is an additional dropdown box to select the year. The selection of an item in the dropdown menu automatically updates the choropleth.

Map

Description automatically generated

Figure Choropleth dashboard

# Crop Trade

This section outlines an analysis performed on the crops in terms of trade. A dataframe for Irish trade in barley, wheat, and potatoes is constructed containing the headings – Import Quantity, Import Value, Export Quantity, Export Value, and Production. Analysing the pairplot in Figure 13, there appears to be a positive correlation between potatoes production and export quantity. There appears to be a negative correlation between production and import quantity. This observation is understandable – the more a country produces, the less it needs to import and more it can export. Interestingly, this effect is not as visible for the other crops. The import data is for the other crops is somewhat stationary when measured against production. This suggests there are other factors at play (production is not directly linked to import/export, production is scaled with demand, etc)

Diagram

Description automatically generated

Figure Pairplot for Irish trade of a combination of barley, wheat, and potatoes between 1961 and 2021

Historical crop price data is now imported to investigate the relationship between production and price for a selection of countries. Price data is acquired from FAO.

Wheat price is plotted versus production for France, Germany, Ireland, Poland, and Spain between the years of 1991 and 2020. It appears that there is no clear correlation between price and production for any of these countries. Prices are not adjusted for inflation which is expected to explain for some of the price variability over the years.

Chart, scatter chart

Description automatically generated

Figure Price vs Production of Wheat (1991-2020)

A k-means clustering model has been implemented to predict a country based on price/production. The model used is the KMeans module by sklearn. The model is fit with the number of clusters given as the number of unique countries in the selected dataset. The model is visually assessed as having 100% accuracy in predicting country from Price/Production. While successful, it must be noted that this particular selection of countries were chosen due to their separation from each other i.e. this is a biased sample selection.

Chart, scatter chart

Description automatically generated

Figure Price vs Production of Wheat (1991-2020) - K-Means

# Sentiment Analysis

## Sentiment

For the topic of sentiment analysis, it was opted to perform a natural language scoring technique on an annual review document published by Ireland’s Department of Agriculture, Food and the Marine. The document chosen for analysis is *Annual Review and Outlook for Agriculture, Food and the Marine 2022*. This document outlines the current state and historical trends of a range of different agriculture and marine topics and is considered to be a appropriate reference for sentiment analysis as it is an extensive and unbiased government document. This is considered a more reliable source of data than other sources such as Twitter and other media outlets that could contain biased actors.

The document itself is available on the XX website in PDF format. In order to interpret a PDF document, the PyPDF2 module was used to read the PDF and convert it to text. It was chosen to split the text into a list of texts, one item for each page of the PDF. This meant that the analysis could be constructed to include only the relevant sections of the document.

After reviewing the table of contents of the document, it was decided to analyse chapter 3.7 only – *Cereals.* Other chapters in the document were not relevant to the chosen topic of crop production.

A function named get\_sentiment\_scores() was written that intakes the pdf address, the start page that the analysis starts from, and the page the analysis finishes. The text was further split into sentences and analysed as such, with a score given to each sentence as scored by Textblob, a natural language analysis tool. The most negative score a sentence can get is -1 and the most positive a sentence can get is +1. A distribution of the allocated scores is presented in Figure 16.

Chart, histogram

Description automatically generated

Figure 16 Sentiment scores for Chapter 3.7: ‘Cereals’ in ‘Annual Review and Outlook for Agriculture, Food and the Marine 2022’.

It is apparent in Figure 16 that there is a positive bias to the scoring in terms of magnitude, indicating that there is positive sentiment in the subchapter. However, there are multiple medium sized bars in the negative territory.

Therefore, it is decided to adopt the common practice of counting the instances of negative, neutral, and positive. This also reduces the effect of outliers in the scoring. For this analysis a score under -0.05 is considered negative, a score between -0.05 and 0.05 is considered neutral, and a score above 0.05 is considered positive. The results are presented in Figure 17. Through this method, it is clear that there is a positive bias to the sentiment, suggesting a positive review or outlook to the cereal production industry in Ireland.

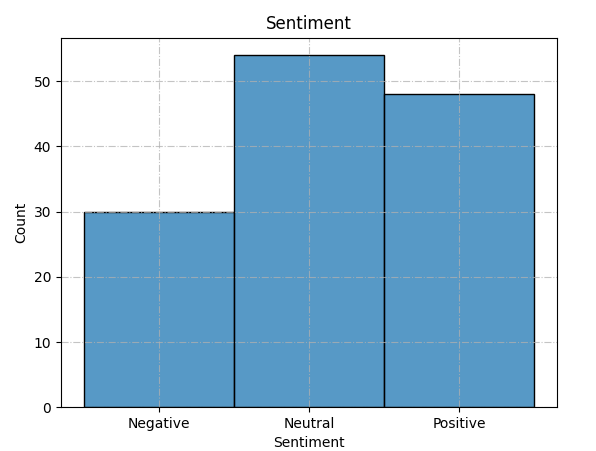


Figure Sentiment counts

## Test using Summarisation with GPT3

In order to test the performance of the analysis above, it was decided to employ a text summarisation technique to perform a quick, manual review. For this, OpenAI’s GPT3 natural language model was used. The model was used within a Python environment using OpenAI’s API. As in the previous section, the PDF was imported using PyPDF and split into pages referring to cereals. Each page was analysed individually using GPT3, with the model given the text prompt “tl;dr”. This prompt requested the model to return a short summarisation of each page. The summary was read by the Author and was confirmed to be positively biased, with most pages containing a particularly positive point about the state and/or outlook of the topic in question. The summary is available in XX

# Conclusion

Crop yield data for the top three Irish crops in terms of production has been acquired and compared against data for other EU countries. Exploratory Data Analysis was utilised to reveal that for these three crops, Ireland had the highest yield in the European Union. A brief investigation was held into the origin of these strong yields. After showing that there is a geospatial factor to crop yields, rainfall was tested and was shown to have a correlation of crop yields. Government spending in agriculture did not reveal a correlation. Multiple inferential statistical tests were run over the course of this study, with hypotheses being put forth and resolved. Multiple attempts at other hypothesis tests were made but not included in the report. This is due to the data being determined as not appropriate for the give test (non-normal, dependant, etc). Difficulty was found in the formulation of valid tests.

Using a linear regression model, future wheat crop yields was determined. GridSearchCV was not utilised due to it not being appropriate for linear regression. Using a K-Means clustering model, specific countries could be identified given their trading price and production quantity of wheat. Using Dash, an interactive dashboard was created showing crop yield in Europe. The interactive element was that the crop type could be chosen, and the choropleth would be repainted. A sentiment analysis was performed on a relevant chapter of a government issued document. The returned sentiment was compared against a summary provided by OpenAI’s GPT3 model.

The data used in this report was acquired and manipulated with ease. As a common source was used for most of the data (FAO), there was a common format for each file. Furthermore, the data was clean with no instances of NaN’s in the timeseries However, there were cases of countries missing from some of the datasets. In hindsight, carrying three crop types through the analysis may have been a mistake. Given the low word count, it could have been better to analyse a single crop in more detail.

# References

Department of Agriculture, Food and the Marine, 2022. *Annual Review and Outlook for Agriculture, Food and the Marine 2022,* Dublin: Department of Agriculture, Food and the Marine.

Food and Agriculture Organization of the United Nations, n.d. *FAO Stat.* [Online]   
Available at: https://www.fao.org/faostat/en/#data  
[Accessed 2 January 2023].

Quantum, 2019. *Data Science project management methodologies.* [Online]   
Available at: https://medium.datadriveninvestor.com/data-science-project-management-methodologies-f6913c6b29eb  
[Accessed 5 January 2023].

Zahrnt, V., 2011. *Food Secutiry and the EU's Common Agricultural Policy: Facts and Fears,* Brussels: European Centre for International Political Economy.

# Appendix

## GPT3 Summary of the Cereal chapter

*“The cereal sector in Ireland is relatively small in comparison to our EU counterparts, accounting for only around 6% of agricultural area. However, tillage is an important cog in the agri-food sector and is a key supplier of high-quality grain to the feed industry. Ireland’s tillage sector produces approximately 30% of the feed materials required by the feed industry to supply the livestock sector. It is increasingly targeting production to the higher-value human food markets, particularly malting and distilling but also the breakfast cereal market. The sector is also an important producer of seed to service its own needs and has a small export market with the potential to increase. It also supplies straw to the livestock sector for feeding and primarily bedding.*

*The area of spring and winter wheat, barley, and oats increased in 2021, while the area of non-cereal crops decreased slightly. Straw production was strong in 2021, and prices ranged between €15-€20 per bale.*

*Cereal production in the EU is forecast to increase by 4.3% in 2021/22, due to favourable weather conditions in many countries. However, some countries did experience drought and excessive cold & wet conditions, resulting in a decrease in production.*

*The tillage sector in the EU is expected to see increased production in the 2021/22 marketing year, due to higher prices and favourable weather conditions. Financial assistance measures such as the Straw Incorporation Measure and the EU Protein Aid Scheme are expected to continue to support the sector.*

*Tillage farmers had a record year in 2021, with family farm incomes reaching €846/hectare. This was due to increased areas of higher-yielding winter crops, as well as higher overall cereal area and increased yields of both winter and spring crops compared to 2020. There was also ample supply of straw to meet demand, despite concerns raised in some quarters. Winter sowings in autumn 2021 for harvest in 2022 are estimated to have remained at normal levels.*

*The tillage sector is an important part of the Irish agricultural industry, contributing significantly to the economy and providing a vital source of food and drink. The sector is facing challenges from the commodity markets, but is diversifying into producing for higher-value markets, which is aiding its viability. From an environmental sustainability point of view, the tillage sector is a low emission farming system, and the draft CAP Strategic Plan proposes a continuation and enhancement of the Protein Aid Scheme to further increase native protein production. There are added advantages of leguminous protein crops as they do not require any form of nitrogenous fertilisers, and the Scheme will help the sector meet ambitious targets for reduction in chemical nitrogenous fertiliser usage by 2030.*

*The tillage sector in Ireland is facing some challenges in the form of Brexit and the loss of plant protection products, but there are also some opportunities for the sector in the form of increased demand for Irish cereals and seed potatoes.*

*The Irish government has announced a €12.2 million tillage intervention package in response to increasing commodity prices and the war in the Ukraine. The package includes a tillage incentive scheme, an expansion of the EU-funded protein aid scheme, and a multi-species sward scheme. There is strong demand for Irish malt and organic oats, and opportunities for expansion in the value-added food sector.”*