WORKSHOP ON DATA COLLECTION AND ANALYSIS FOR CODEX PROCEEDINGS

Data Analysis for the Development of Maximum Levels (MLs) – Part 1

Dr. Amine Kassouf, GFoRSS

Outline

1. Introduction
2. Data Selection & Extraction
3. Data Clean-up
4. Generating Overview Data
5. Conclusion

GFORSS
Global Food Regulatory Science Society

Training Objectives

By the end of this training module, participants will be able to:

1. Describe the step-by-step procedure for setting Maximum Levels (MLs)
2. Identify and extract relevant data for ML establishment
3. Understand and apply basic data clean-up techniques
4. Generate initial summaries and overviews of the collected data

Introduction

DEVELOPMENT OF MAXIMUM LEVELS

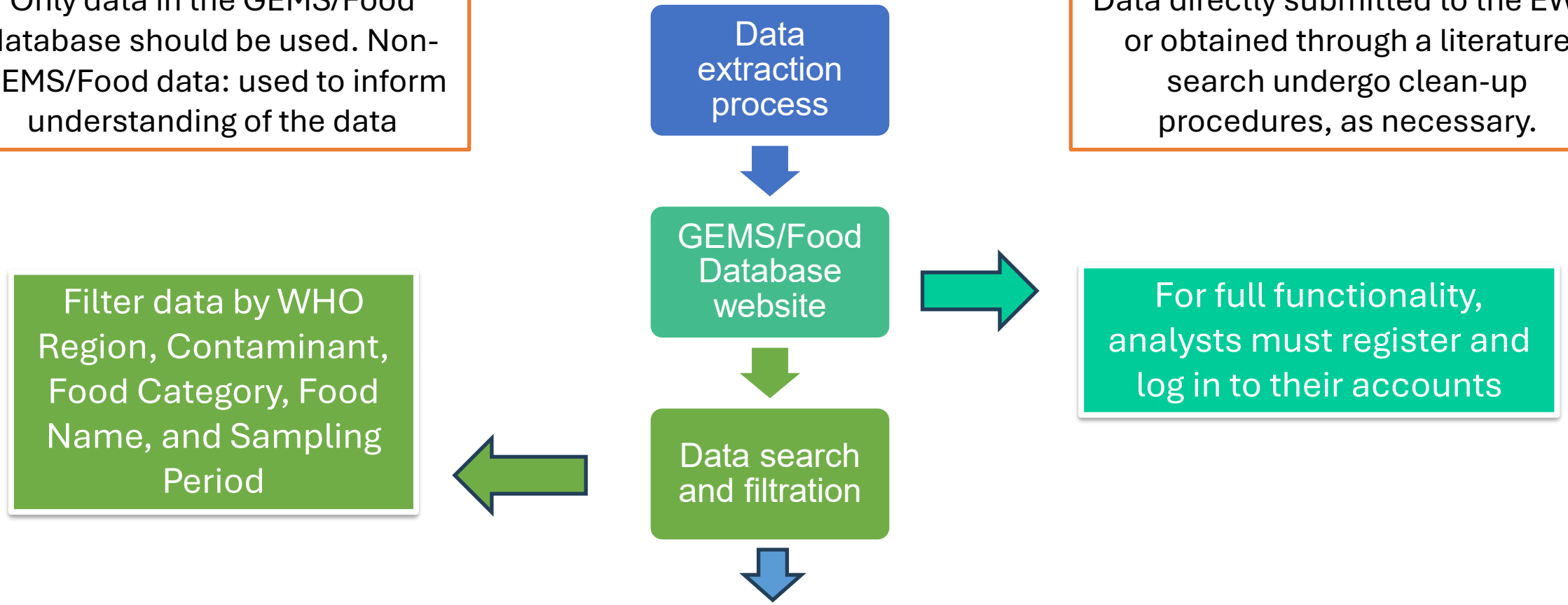
1. **Data Extraction**
   - To identify the dataset for ML development
2. **Data Selection / Clean-up**
   - To support consistent and transparent handling of data across different sources
3. **Data Analysis: Generating Overview of Data**
   - Descriptive data exploration
4. **Statistical Analysis of Occurrence Data**
   - Analyzing the distribution of the dataset
   - Handling of Datasets
   - Data illustration
   - Descriptive statistics
5. **Calculation of Rejection Rates**
   - At hypothetical MLs
6. **Calculations of the reduction of dietary exposure**
   - To the contaminant from the target commodity at hypothetical MLs
7. **Final Recommendations to CCCF and Report Preparation**

GFRSS
Global Food Regulatory Science Society

Data Selection & Extraction

Only data in the GEMS/Food database should be used. Non-GEMS/Food data: used to inform understanding of the data

Data directly submitted to the EWG or obtained through a literature search undergo clean-up procedures, as necessary.



- ❑ These filters will allow the analyst to identify data respectively to a particular Call for Data or ToR.
- ❑ Best to consult with the GEMS/Food database administrator/ Confidential data.

GFORSS
Global Food Regulatory Science Society

Which of the following statements is TRUE regarding the use of data for ML setting in Codex work?

A. All data from any national database can be used without verification

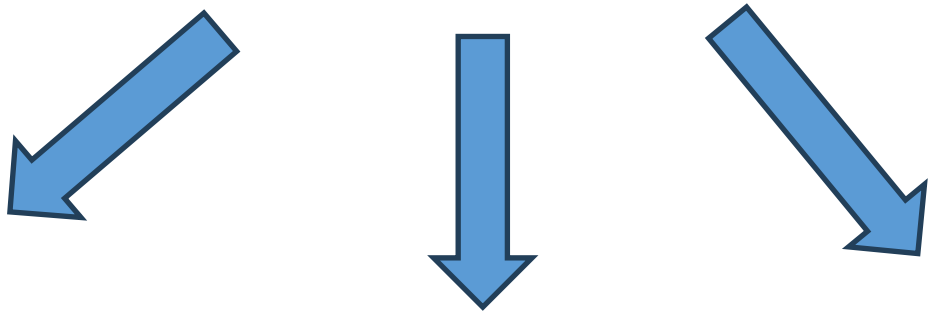
B. Only non-GEMS/Food data are used to derive MLs

C. GEMS/Food database data are preferred; non-GEMS data may be used to support understanding

D. Literature data should always replace GEMS/Food data

Data Clean-up

Purpose: to remove non-relevant or inappropriate samples from the dataset before analyzing the dataset to recommend MLs



Samples that are clearly outside the ToR
e.g. ketchup samples for a work on tomato sauce

Samples that are outside the date range of the ToR
(the case of adoption of a CoP!)

- Samples with unacceptably high LOQs
- Accuracy of the data cannot be confirmed, and corrections cannot be made

Samples missing crucial information:
- sample information
- the state of the food analyzed
- inadequate local food identifier (e.g. for MLs derived for species)

Data Clean-up

Data for which the information on the portion analyzed is not clear (e.g. peeled vs whole fruit, or husked rice vs polished rice)

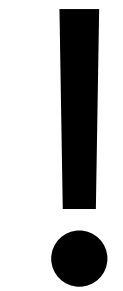
- Contact the point of contact for the country/organization or observer that submitted the data for clarification.
- Or assume that the portion was analyzed in the state that it is usually sold/consumed.

If no reasonable assumption can be made, **these data should be excluded**

GFRSS
Global Food Regulatory Science Society

Data Clean-up

Data originating from suspected fraudulent/economically adulterated samples



Possible signs of fraudulent/economically adulterated samples are:

- certain samples are orders of magnitude higher than others, e.g. 0.1 mg/kg versus 100 mg/kg, or
- temporal variability in data (depending on contaminant), e.g. data are much higher in one year of the dataset.

Mycotoxins vs Lead?

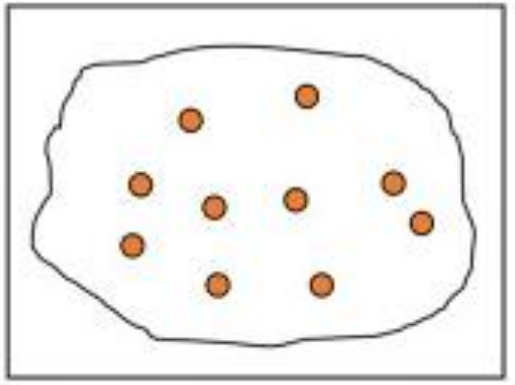
Data that are clearly related to fraudulent/economically adulterated samples should be excluded from the analysis and the exclusion must be documented.

GFRSS Global Food Regulatory Science Society

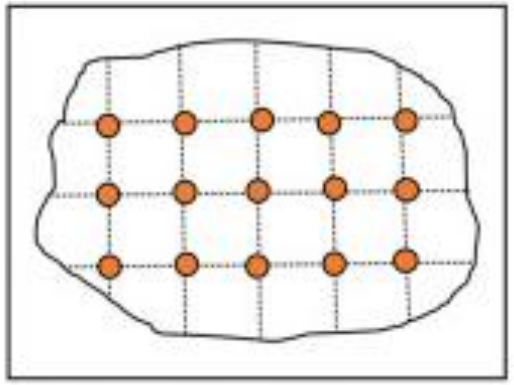
Data Clean-up

Data from targeted sampling and bias

- Targeted sampling data should not be used in the derivation of MLs as they are not representative of the general population and may not reflect achievable levels in regular situations.
- Some bias could be introduced in random sampling: however still reflect realistic variation in the occurrence data.



Random sampling



Systematic grid sampling

Data Clean-up

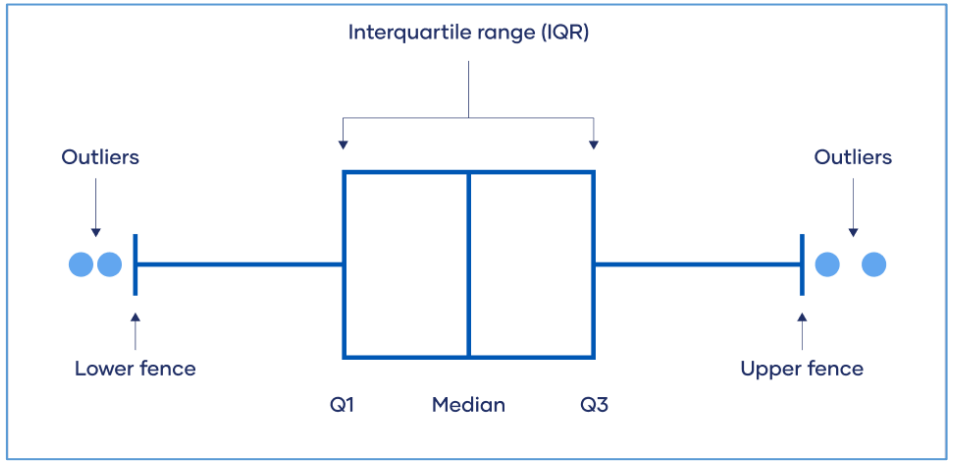
Determination and handling of outliers/extreme values

General Rule

Outliers/extreme values not be discarded unless there is a valid reason to do so!!

Extreme values can be valid values:

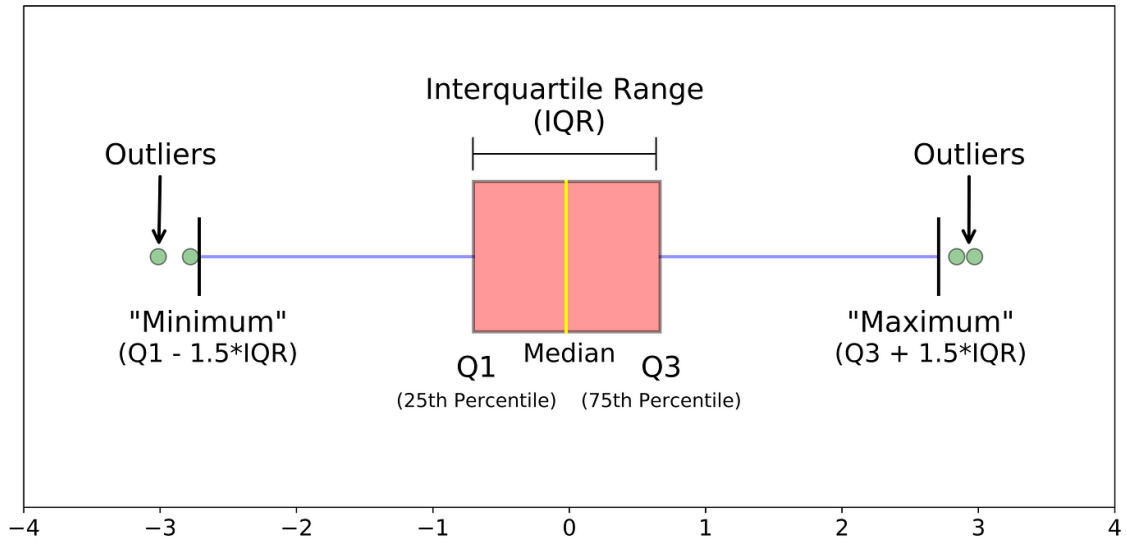
- Heterogeneity of contaminant distribution (such as hotspots for mycotoxins) or due to natural variation of measured contaminants (e.g. resulting from weather conditions, soil condition, etc.).
- Erroneous, e.g. errors in measuring and processing data, including incorrect calculation or using the incorrect unit of measurement,
- Result from fraudulent behavior (economic adulteration).



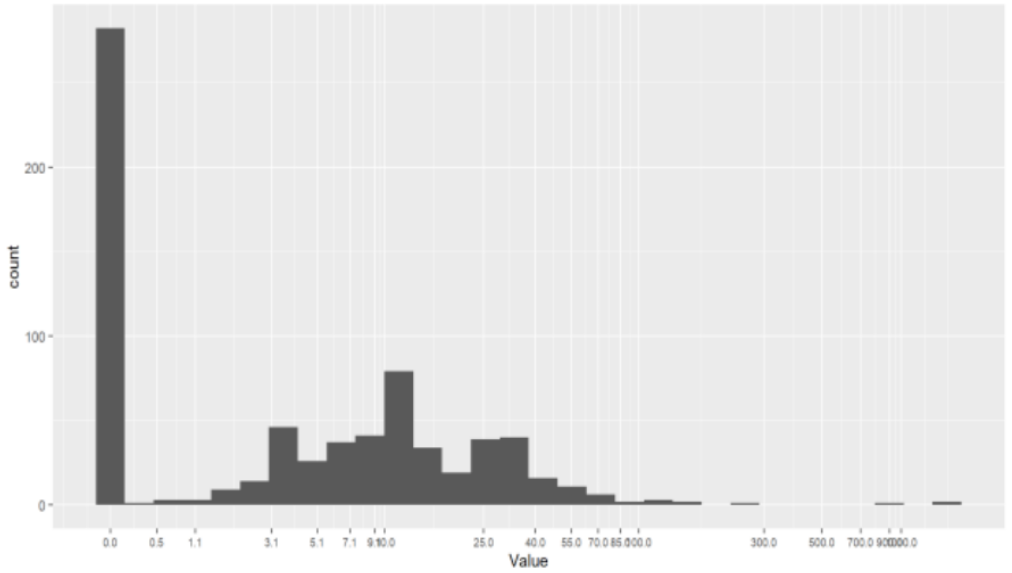
GFORSS
Global Food Regulatory Science Society

Data Clean-up

- High percentile values, rather than maximum values, are used as a basis for the development of MLs based on rejection rates, the impact of outliers on derived MLs will usually be small.
- In cases where a notable percentage of data points (e.g. 2–5%) are excluded from the dataset as outliers, this could affect the high percentile values.



The interquartile (IQR) approach



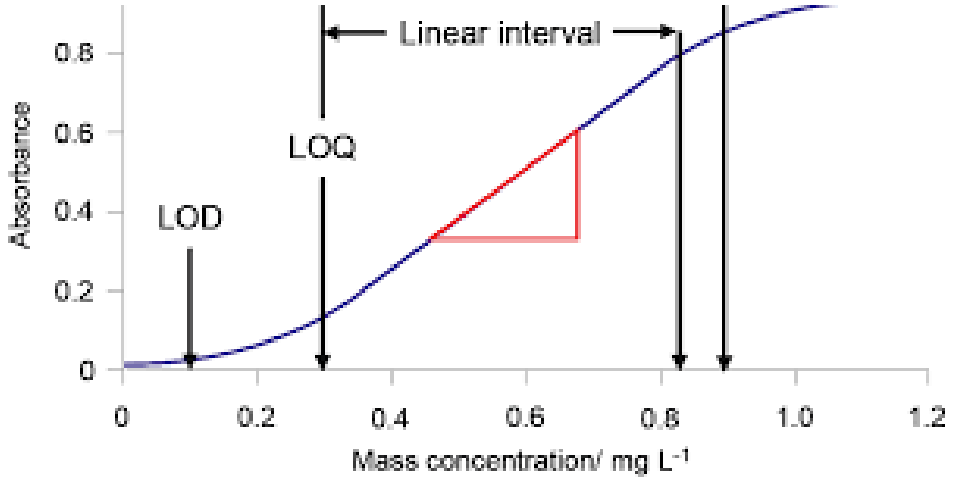
The visual approach (not recommended)

GFORSS
Global Food Regulatory Science Society

Data Clean-up

Guidance for data inclusion/exclusion in different LOQ/LOD scenarios

- **In the case where no LOQ/LOD is provided for a specific dataset:**
  - The submitting country/organization or observer should be contacted as a first step to obtain such information (i.e. LOD and/or LOQ).
  - In the case where the dataset contains (nearly) only quantified results: the data set could be used.
  - In the case where the dataset contains a significant amount of left-censored data (individual data without quantified (finite) values, generally referred to as data below the reported LOQs/LODs): data set should not be used.
- **In the case where an LOQ is provided:**
  - Identify a cut-off level for the LOQ depending on the MLs being considered (e.g., LOQ < ML under discussion, LOQ < 1/2 ML under discussion).
  - The dataset contains a significant amount of left-censored data: further contextual information should be considered (e.g., are data from an importing country or producing country).



GFORSS
Global Food Regulatory Science Society

True or False:

Data that appear as outliers or extreme values should always be excluded from the dataset during clean-up.

Why should data from targeted sampling generally not be used in setting MLs?

A. They are too expensive to analyze
B. They overestimate typical exposure and are not representative
C. They include only processed foods
D. They contain duplicated entries

Generating Overview Data

Overview of countries, number of data points, period of data coverage

Country of production			
Production/harvest year			
Amount of data included and excluded			
…			
….			

All steps taken in the data clean-up and the rationale and assumptions made should be provided.

Useful to provide information (e.g. from FAO) on the major production regions for the commodity under discussion

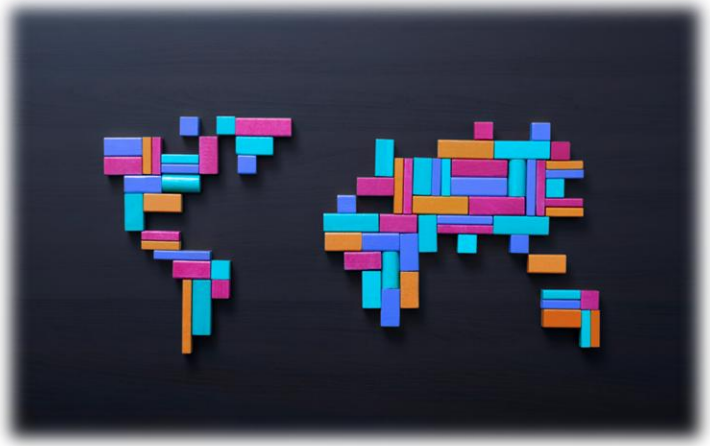
More focused analysis of specific geographical areas and time periods.

GFORSS Global Food Regulatory Science Society

Generating Overview Data

Geographical coverage of the provided occurrence data

- ❑ When evaluating an ML for a particular commodity, the dataset should include representation of **production regions** that are important to **international trade.**
- ❑ **Data from importing countries** might be biased if the food has to comply with their requirements such as a previously established ML.



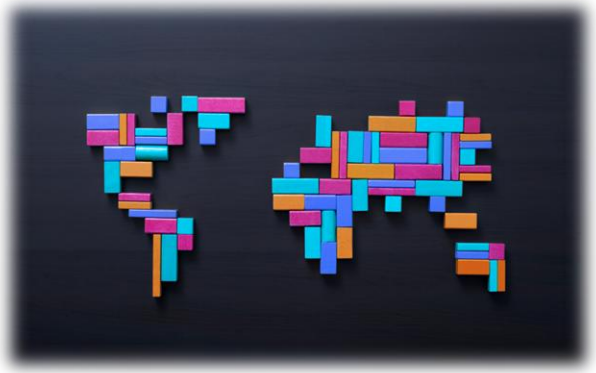
- Additional contamination could take place during transport from the producing country (e.g. mycotoxin production)!
- Provide separate analyses for producing and importing country datasets?

GFORSS
Global Food Regulatory Science Society

Generating Overview Data

Geographical coverage of the provided occurrence data

- Only if there are enough data that show an indication of large differences in reported levels between regions or between countries in a region, analysis could be performed by region or country.



No geographical representativeness

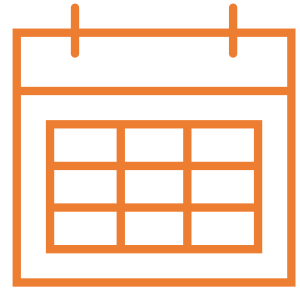
Compromise: If there is an urgent need to establish an ML for consumer health protection, a compromise needs to be reached in CCCF to set an ML based on available data.

In these cases, the ML can be reviewed within 3–5 years to assess whether adjustments are necessary when more data are available.

Generating Overview Data

Time period coverage of the provided occurrence data

- ❑ Year-to-year variation or evolution of contamination in time e.g. mycotoxins where data from the last 10 years may provide a very good representation of the year-to-year variation
- ❑ It may be appropriate in certain cases to perform a time trend analysis.



If a Codex CoP has been established and implemented, the data under consideration should be from the years after the implementation of that CoP to reflect good agricultural and manufacturing practices.

True or False:

If occurrence data lack full geographical representativeness, Codex cannot proceed with setting an ML under any circumstances.

Conclusion

The process of setting Maximum Levels (MLs) relies on a structured, stepwise approach grounded in data.

Accurate **data selection**, **extraction**, and **clean-up** are critical to ensure reliability and transparency.

Initial **data exploration** helps guide further analysis and decision-making.

GFRSS
Global Food Regulatory Science Society

ARAB CODEX INITIATIVE
المبادرة العربية للدستور الغذائي