EnviroInfo 2009 (Berlin)
Environmental Informatics and Industrial Environmental Protection: Concepts, Methods and Tools
 ISBN: 978-3-8322-8397-1

Drilling Down Multiple Data Sources for Risk Assessment and Cost Benefit Analyses: The Story of a "Tomato Mashup" for an Integrated Approach.

Cristina Ford McLaughlin,
US Food and Drug Administration
Center for Food Safety and Applied Nutrition
College Park, MD 20740
Cristina.McLaughlin@fda.hhs.gov

Abstract

Public health agencies responsible for food safety are facing many information challenges as food production and distribution systems, become increasingly complex. Integrated risk assessments and cost-benefit analyses can help decision makers choose appropriate policies to address major public health problems. One component of an integrated analysis is the exposure assessment. The exposure assessment is the probability of exposure to a particular food associated with a particular hazard or nutrient. The probability of exposure can be characterized by integrating relevant supply chain information in a model. However, as terabytes of information are being generated for multiple purposes and by diverse entities and institutions, analysts are often encumbered by efforts in finding relevant data or deciding how to choose the best data for their analyses. Thanks to developments in informatics, some of the tools used by analysts such as decision support technologies can better highlight the merits that a model can achieve because they allow integrating multiple data sources, visualizing the data –geographically, as an influence diagram or other form of graphical expression-- and also provide an interface that prevents the separation of end users from analysts. This is better illustrated by following a tomato through different points in the supply chain. The data from different sources can be made relevant in estimating both probability -based and frequency-based consumption and exposure to potential hazards or nutrients.

1. Introduction

As global populations and income levels increase, global demand for food that is safe and nutritious is also increasing. As a result, more food producers, processors and distributors worldwide are attempting to meet this growing global demand for food by depending on an ever expanding and complex global network. The increase in size and complexity of food production and distribution systems in the U.S. has led to lower food prices and greater variety of foods (Craig 2006).

An increase in global food trade can introduce new food risks or increase existing food risks. Increasing risks in food safety may require more complex regulatory systems which in turn increase the demand for research on food safety issues. Increasingly complex regulatory systems create new information challenges for health authorities. The effectiveness of policy decisions is highly dependent on the quality of information analysts have to inform policy makers in their decisions. As an ever increasing amount of information is collected, managed and disseminated by a number of diverse entities, the need to convert this information to practical use is more challenging.

2. Risk Analysis as a Cognate of Informatics

Informatics and Risk Analysis are both multidisciplinary fields with multiple dimensions. According to the definition in Wikipedia, "*Informatics studies the science, practice and the engineering of information*

systems. It studies the structure, algorithms, behavior, and interactions of natural and artificial systems that store, process, access and communicate information. It also develops its own conceptual and theoretical foundations and utilizes foundations developed in other fields. Since the advent of computers, individuals and organizations increasingly process information digitally. This has led to the study of informatics that has computational, cognitive and social aspects, including study of the social impact of information technologies".[1] In other words, informatics can be defined as a multidisciplinary and multidimensional use of technology to support the discovery of knowledge dissemination decision makers need to achieve their policy objectives.

Risk analysis according to the standard paradigm is made up of three components: risk assessment, risk management and risk communication. It is also argued that components should be kept separate in the interest of scientific integrity and to make sure results are not affected by political pressure. However, the interaction among the different components of risk analysis is necessary so that the risk assessment along with the CBA can aid risk management in decision making. Thus, a more complete definition for risk analysis states that it is the framework for decision makers and regulators to use risk assessment, risk management, cost-benefit analysis (CBA), and risk communication in order to reduce the public impact of risks to health, safety and the environment.

Risk assessment is a process, like CBA, that can help risk managers decide whether a potential hazard is of enough significance that it needs to be managed or regulated. Thus, when conducting such analyses, some level of risk is considered acceptable. Risk Assessment has four components that are parallel to components of CBA: 1) *Hazard identification* is equivalent to impact identification which implies that either the markets or the government failed and that there is a need for regulation (or de regulation); 2) *Dose Response* which deals with the relationship between specific quantities of something that is causing a health effect or illness. 3) *Exposure assessment* which can be attributed to a particular behaviour or characteristic in one segment of the industry or a subset of the population; and 4) *Risk Characterization* is very important to economists because valuation of benefits depends highly on risk characterization aspects such as severity. For example the medical costs of an illness will vary according to the different levels of severity in a disease. Table 1 illustrates how components of a risk assessment relate to components of a cost benefit analysis.

Risk Assessment	**Cost Benefit Analysis**
Hazard Identification	Impact Identification (pathogens in food)
Dose Response	Economic Cause and Effect (Industry Practices and Consumer Behavior)
Exposure	Specific Cause (producer and supply chain characteristics)
Risk Characterization	Economic Characterization (cost of illness by severity)

As part of a risk analysis framework, an integrated risk assessment and cost benefit analysis not only helps policy makers make informed decisions, but it is also useful in characterizing uncertainty and highlighting knowledge gaps. Measuring the benefits of reducing the population risk of illness requires inferring the value of prevention from limited or unwieldy information about a contaminated food. A general simplified

[1] http:www//wikipedia.org. June, 2009

model to measure benefits involves a five-part calculation, using the following factors: (1) The quantity of contaminated food sold or consumed; (2) the per unit contamination level (i.e. number of contaminated servings); (3) the probability of illness; (4) the number and severity of illnesses prevented; and (5) estimated health benefits in dollars. The data sources needed to quantify these five factors come in multiple types from multiple sources, disciplines and institutions and in varying levels of quality. The data may have originally been collected to answer different policy questions over different time periods. It is incumbent upon analysts to choose the data according to their own judgment regarding relevance and quality. Advances in Bayesian approaches used in risk assessment are leading to the creation of more data and models that can be integrated in an economic analysis. Economic theory provides methods to characterize trade-offs. This interaction of disciplines calls for a common information taxonomy that can facilitate the successful integration of analyses by the successful integration of data.

3. Information Challenges in Food Safety Risk Analysis

Decision makers rely on a risk analysis framework to evaluate risk management options for implementing food safety programs and policies. However, the effectiveness of food safety policies depends highly on the quality of information they have to make decisions. A recent Food Safety Research Consortium (FSRC) report, by (Taylor and Batz 2008), addresses the many scientific, technical, legal, regulatory and industry issues and how they affect how food safety data are collected, managed and disseminated in the United States. For example the information flow generated by a food borne outbreak investigation can be valuable to the development of a long term prevention program. Unfortunately the results of many outbreak investigations are incomplete because of resources and time constraints. In contrast the information flow in a microbial risk analysis decision describes a wide array of potential data from many sources and is designed in a slower more deliberative process (Taylor and Batz 2008).

Increase in global food trade can cause an increase in food safety risks, or introduce new risks, which in turn require increasing complexity in food regulatory systems to address the risks. Addressing new complexities require introducing new variables and even dimensions to existing information. Finding relevant data or information for this multi-disciplinary effort is difficult because "knowing what to find" requires expert knowledge of a subject. As the authors in Gammack, et al 2006, have eloquently stated: -*"In search engines, relevance is key to finding what you want, but like beauty, relevant is in the eye of the beholder. Search engine use is not going to help find this implicit type knowledge or about knowledge that has not yet been articulated into words or other representations. This is where knowledge elicitation and discovery comes in. The elicited knowledge usually comes from a human informant in many forms, medical interviews, job interviews, chat show interviews, witness questioning in court, speed dating, and everyday life."*(Gammack et al 2006)

Thus, as amounts of information are growing at a formidable rate, so are the challenges and opportunities in finding and making good use of relevant data. Integrating a risk assessment into a cost benefit analysis often involves using heterogeneous data sources in order to address policy questions. Answers to most policy questions are never simple and are riddled with variability and uncertainty. Consequently, risk managers might make decisions with uncertain or incomplete information on a different basis than the one recommended in the economic analysis. In fact, depending on how much uncertainty is characterized in a decision, similar analyses may play a differing role in the decision outcome (Williams and Thompson 2004).

This also may bring new challenges in addressing complexity in a manner that is intuitive and transparent.

Other challenges include addressing data compatibility, especially with geographic data. Different players in the food safety arena need unencumbered information exchange capabilities such as applica-

tions that allow multiple collaborators. Through such applications, the relationships among the basic categories and entities that make up the organizational framework for food safety risk analyses can be identified, so that key content can be accessed more quickly, and the bases of the technical analyses can be more readily understood. Information and communication technologies present significant opportunities for better integrating risk assessment and CBA to guide policies for food safety, including strengthening information access and sharing knowledge that supports these analyses and to further the understanding and implementation of effective decisions for public health protection.

Current developments in information and communication technologies (ICT), especially in programmable web 2.0/3.0 applications address the need for combining multiple data sources in the following section.

4. From Web 2.0 Mashups to Semantic Web Smashups and other Decision Support applications.

The term "Mashup" was used originally to describe mixing together different musical tracks and create a new musical piece. A web mashup is a web application made from combining content, presentation or application functionality from disparate web sources. Types of mashups include consumer mashups that emphasize presentation such as Google Maps and enterprise or data mashups that combine data from internal and external sources. Examples of mashups can be found at http://www.programmableweb.com/

Developments in information and communication technology (ICT) can improve effectiveness, quality and transparency in many areas. Developments in programmable web and Semantic Web (web 2.0/3.0) once thought of as competing visions in the evolving web are now suggesting that their core technologies are more complementary. The community and usability of Web 2.0 coupled with the infrastructure of the Semantic Web is more apt for mashup like information sharing (Ankolekar et al 2007). Smashup is a new term for describing a semantic mashup.

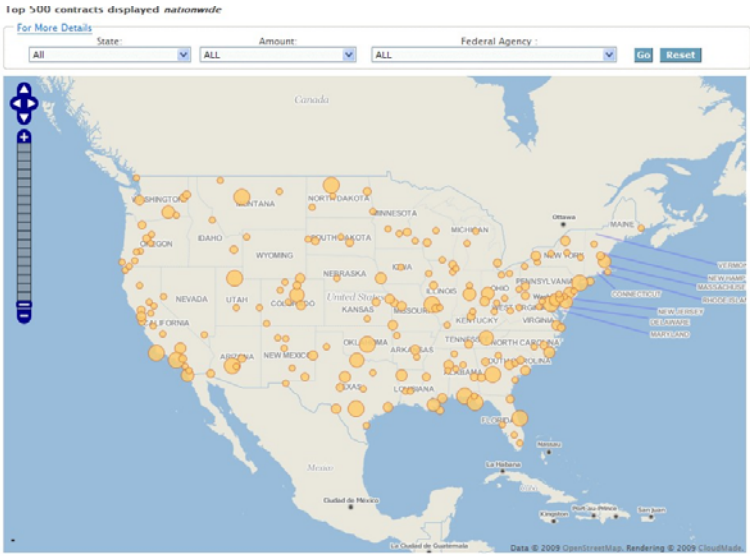
There are many cognates in the field of informatics. However, since the completion of a growing number of major genome sequencing projects, the biology cognate otherwise known as bioinformatics has been transformed. With a goal of sequencing three billion bases, the only limitation was how fast the technology progressed and how much funding the projects received (Rojas et al 2009). Bioinformatics type research with genomic datasets would typically involve downloading the data and application program interfaces (APIs) from a small number of institutions, like the National Center for Biotechnology Information (NCBI). Researchers would then combine pieces of data from different sources into a new piece of data. In the earlier literature, bioinformaticians didn't describe this type of processing as a mashup, but conceptually they sound very similar. With the increasing popularity of the semantic web technology and a growing number of databases in bioinformatics, researchers in Canada, developed BIO2RDF which is a system to integrate bioinformatics knowledge into a mashup (Belleau et al 2008).

The Semantic web (web 3.0) is beginning to move beyond the earlier adopted bioinformatics to more broader chemical and physical sciences and also in integrating the work of smaller scale operations producing more diverse data (Frey 2009). More and more mashups in environmental and public health communities are also being created and integrating into subspecialty areas such as epidemiology and Public Health Geography. Semantic web is also facilitating powerful data integration from heterogeneous sources to having a flexible way to compose a semantic mashup within multidisciplinary fields such as health care management and life sciences (Cheung et al 2008).

This type of information sharing can help analysts build models that are more transparent and help recognize and communicate information and knowledge gaps early in the policy formulation process. Similar application interfaces, such as Enterprise Mashup, Business Intelligence (BI) and Decision Support Systems (DSS) are offered in simple to build forms that can be used by businesses and institutions. Such applications can offer more intuitive appeal because their building architectures are not only targeted to

smaller audiences but they allow for combining data from different sources and allow communication across multiple entities, stakeholder, and disciplines(Ogrinz 2009). More importantly, they can translate data into visual media such as maps and charts. Decision makers and analysts alike can explore complex models starting with a map, influence diagram or a chart and drill down the data sources or reports for more detail and create new ways of representing data, diagrams or maps.

An innovative example of a mashup can be found at https://www.Recovery.gov. The Recovery Act was enacted to create and save jobs through projects funded at the state and local level. States recently began to send information to federal agencies with details on how they are spending some of the funds. The Recovery Board regularly updates the information displayed on this map, as more data is received from the States. [1]



5. Insights for Composing a Tomato Supply Chain Mashup.

Using a desktop version of a DS application (Analytica™) we build an influence diagram that integrates different data sources for building --at least conceptually--components for a data mashup of the tomato supply chain. Each source of data is characterized as a probability distribution that feeds into the model. This particular type of application allows for complex information to be compressed into simpler presentations for analysts and end users alike. This not only allows for a more creative integration of different types of data from different sources but it adds transparency to the analysis without compromising data quality. The integration of multiple sources can help meet legal requirements calling for analyses of policies involving food safety; it can also lead to better analyses.

As an example of integrating data from more than one source, consider the multiple consumption pathways taken by fresh and processed tomatoes consumed at home or away from home. Following a tomato throughout its supply chain shows how different data sources can be combined into composing a mashup that presents both probability-based and frequency-based estimates of consumption and exposure to hazards or nutrients that are associated with practices in a particular point in the supply chain.

Using a simple hypothetical example of salmonella contamination in tomatoes, the social benefits of reducing the risk of illness can be estimated. The probability of illness due to contamination of a tomato

[1] https://www.Recovery.gov. July10, 2009

shipment can be associated with tomatoes that come from a particular place (point of production risk) or it can be associated with a particular population cohort such as the elderly in a nursing home.

Chart 1 shows different pathways tomatoes take on their way to consumers. In this framework, exposure or intake would be a function of the characteristics of the different pathways from the grower, shipper, packer, processor, or importer to the retailer, restaurant, or home. Some or all the different variables described in Chart 1 can be combined in an exposure assessment.

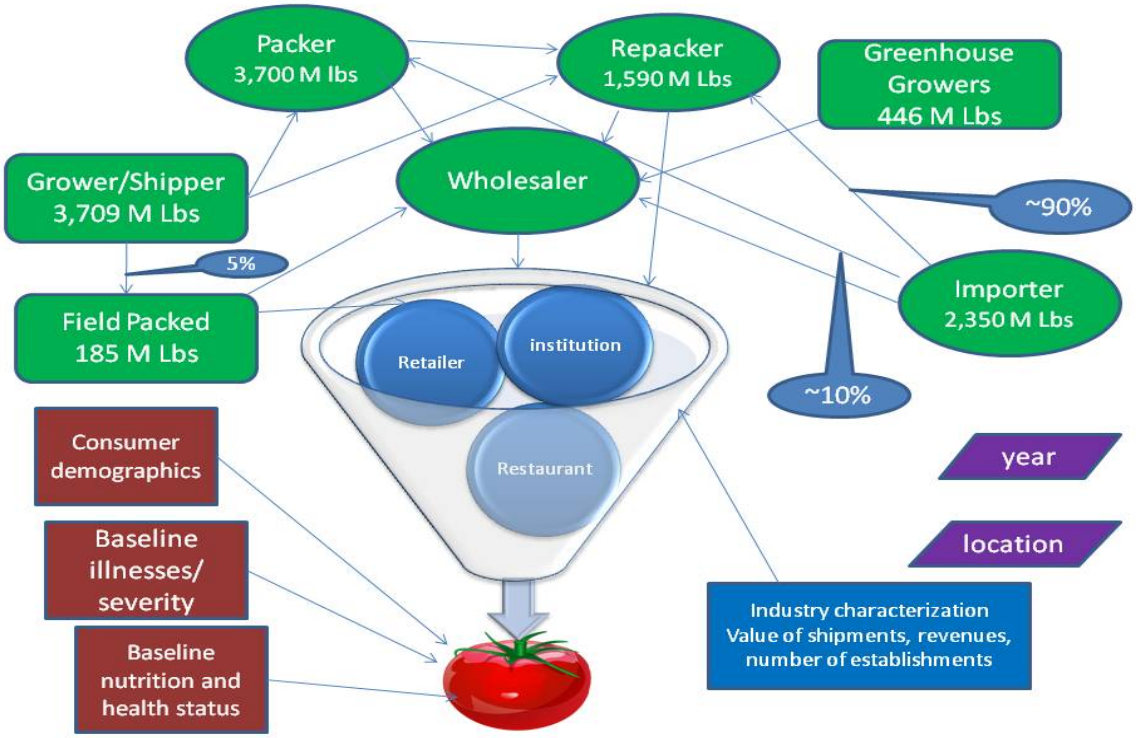


Chart 1. Fresh Tomato Pathways From Grower to Consumer[1]

6. Examples of Data Sources Used in Estimating Exposure.

The following list of data sources provides their descriptions with a brief explanation in how they could be used in estimating exposure. Some may be better suited for estimating frequency and others for assigning probability form drawing inferences about our state of knowledge. More sources and their descriptions can be found at: http://www.foodrisk.org.

Government Data Sources

FDA Total Diet Study (TDS)

The Total Diet Study (TDS) collects and analyzes a market basket containing 280 different foods four times each year to determine levels of contaminants and nutrients. The program analyzes the foods for elements, pesticide residues, industrial chemicals, radionuclides, and selected vitamins such as folic acid. The TDS program purpose is to determine prevailing levels of contaminants, not to enforce tolerances or

[1] Poundage estimated using data from http://usda.mannlib.cornell.edu/MannUsda/viewDocumentInfo.do?documentID=1210

other regulatory limits. Because the program is for surveillance, not compliance, the levels of the substances that are measured in the TDS are generally much lower than the levels set by regulatory programs.
Availability: http://www.cfsan.fda.gov/~comm/tds-toc.html

FDA Office of Regulatory Affairs Import Program
The OASIS (Operational and Administration System for Import Support) system provides monthly detention reports on imported products that appear to violate the food and drug laws enforced by the FDA. Although OASIS data captures a small fraction of imports and is not representative, the information from these reports could be used as one of many sources used to estimate the likelihood of consuming violative food products in the absence of import inspections. For example, inferences from OASIS records can be used in calculating the probability that a certain quantity of food is contaminated by using the inferences as previously assigned probabilities according to the rules of probability theory (Jaynes, E. 2003). However, this requires making some assumption about, the proportion of the lot contaminated[1] and how the likelihood of contamination may change depending on the type of food and its subsequent processing or handling steps.

Production: Processing and Fresh Tomatoes U.S. and by State
The USDA's Economic Research Service maintains comprehensive databases on food and agricultural commodities. For example, the U.S. Tomato Statistics (Updated 11/2007- Stock 92010) provides 114 time-series in downloadable spreadsheet files describing fresh and processing tomato markets, including acreage, yield, production, price, value, trade, and per capita use. The database includes state-level production series.
Availability: http://usda.mannlib.cornell.edu/MannUsda/viewDocumentInfo.do?documentID=1210

Commercial Data Sources

Scanner Data
Supermarket scanner data provide very specific point of sale data for food stores, drug stores and mass merchandisers. Detail on individual food items such as prices, units sold and packaging information is collected on a nationally representative sample of stores. This source is extremely useful for estimating the probability of consuming food products in different forms. For example, supermarket scanner data can be used to estimate the relative frequencies of the consumption of fresh, shelf-stable, and frozen fruit and vegetable juices.

Produce credit reporting services
These types of services for the fresh produce industry provide lists of the number of businesses by annual throughput in hundred-weight and by type of operation, such as packers, re-packers, importers, growers, or shippers. This type of information is used to characterize the frequency of each type of operation in the fresh fruit and vegetable supply chain.

7. Conclusion

Estimating food consumption and exposure can start with a multiple pathway model that uses many sources of information and data. Because different data sources have different strengths and weaknesses, analysts may use more than one source to fill in information gaps or to corroborate previous estimates.

[1] One way to estimate the proportion contaminated is to assign the probability of contamination as the proportion that maximizes the probability that one sample k of n composite samples of size c will be positive for the contaminant.

Many sources provide indirect evidence that can be combined with direct evidence to generate estimates of food consumption or exposure to hazards and nutrients.

The data sources described are only an example from many others that can be used to estimate consumption in cost-benefit analyses and risk assessments. For particular problems, other sources might provide necessary data or information. Whatever the problem, however, many exposure analyses based on consumption estimates can be improved by using multiple pathway models that allow comprehensive and imaginative uses of data from many sources. Developments in information and communication technologies (ICT), especially in programmable web applications (Web 2.0) intersecting with Semantic Web could allow users to combine any multiple data set of their choosing, repurpose it with another data set and present it in a new view.

At the end of the day, risk assessors and economists may need to learn web development skills as we once learned how to use common Bayesian techniques in order to fully exploit the data potential. Computer and information scientists may need to become specialized within a particular cognate or multiple cognates in order to meet institutional or disciplinary needs regarding information infrastructure.

References

Henry, Craig, W. (2006) Food Products Association. The U.S. Food Supply: How Changing Demographics and Consumer Demand Pose New Challenges for Food Safety, Addressing Foodborne Threats to Health: Policies, Practices, and Global Coordination, Workshop Summary. Chapter 1 The US Food System, page 43. National Academies Press, Washington, DC.

http://www.nap.edu/catalog/11745.html

Taylor M. and Batz M. (2008): Harnessing Knowledge to Ensure Food Safety: Opportunities to Improve the Nation's Food Safety Information Infrastructure. Food Safety Research Consortium (FSRC Report 08-01). FSRC, Gainesville, Fl and School of Public Health and Health Services, The George Washington University, Washington, DC.

http://thefsrc.org/FSII/

Gammack J, Hobbs V, Pigott D, (2006): The Book of Informatics, Cengage Learning Australia (p. 101).

Williams, R.A. and Thompson, K.M. (2004): Integrated Analysis: Combining Risk and Economic Assessments While Preserving the Separation of Powers, Risk Analysis, 24(6):1613-1623.

Ankolekar A., Krotzsch, M., Tran T. and Vrandecic, D.(2008): The Two Cultures: Mashing up Web 2.0 and the Semantic Web, Web Semantics: Science, Services and Agents on the World Wide Web, Semantic Web and Web 2.0, 6(1): 70-75.

Rojas, I., Pomares, H., Velenzuela, O., Bernier, J.L.. (2009): Applications in Bio-informatics and Biomedical Engineering, Bio-Inspired Systems: Computational and Ambient Intelligence, Springer Berlin/Heidelberg. Volume 5517 pages 820-828.

Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. (2008): Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. J Biomed Inform, 41(5):706-16.

Frey, J.G., (2009): The value of the Semantic Web in the laboratory, Drug Discovery Today, 4(11-12): 552-561.

Cheung, K., Yip, K., Townsend, P., Scotch, M. (2008): HCLS 2.0/3.0: Health care and life sciences data mashup using Web 2.0/3.0, Journal of Biomedical Informatics, 41(5): 694-704.

Ogrinz, M. (2009): Mashup Patterns: Designs and Examples for the Modern Enterprise. 1. Addison-Wesley Professional.

Jaynes, ET. Probability Theory: the Logic of Science, (2003) Cambridge University Press. Page 100.