Check for updates

RESEARCH NOTE

A structured process to create datasets with nutritional information [version 1; peer review: 2 not approved]

Francisco Rodriguez-Montúfar , Bayron Ordoñez-Buitron, Diego Duran, Angela Chantre-Astaiza, Gustavo Ramirez-Gonzalez

University of Cauca, Popayán, Cauca, Colombia

V1 First published: 25 Jan 2018, 7:110
<https://doi.org/10.12688/f1000research.12979.1>
 Latest published: 25 Jan 2018, 7:110
<https://doi.org/10.12688/f1000research.12979.1>

Abstract

There is a lack of datasets in Colombia that characterize the nutritional components and other similar information about food items. This study describes a structured process to develop datasets that captures the preferences and purchases of food items by a selected group of people. The datasets would classify products according to their sodium and sugar content. The outcome of this structured process would include three datasets, each with a different focus: the first contains data on food preferences, the second contains the purchase history according to the invoices obtained, and the third contains characteristics of the food items such as its brand, category, sodium and sugar content levels, among others.

Keywords

dataset, sodium, sugar, purchase invoices, survey

Open Peer Review

Approval Status **XX**

	1	2
version 1	X	X
25 Jan 2018	view	view

1. **Georgina Gómez Salas**, Universidad de Costa Rica, San José, Costa Rica

Yadira Cortes Sanabria, Pontificia Universidad Javeriana, Bogotá, Colombia

2. **Irina Kovalskys**, Autonomous City of Buenos Aires, Buenos Aires, Argentina

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Francisco Rodríguez-Montúfar (jfrodriguez@unicauca.edu.co), Bayron Ordoñez-Buitron (bjbuitron@unicauca.edu.co), Diego Duran (dduran@unicauca.edu.co), Angela Chantre-Astaiza (achantre@unicauca.edu.co), Gustavo Ramirez-Gonzalez (gramirez@unicauca.edu.co)

Author roles: **Rodríguez-Montúfar F:** Investigation, Visualization, Writing – Original Draft Preparation; **Ordoñez-Buitron B:** Investigation, Writing – Original Draft Preparation; **Duran D:** Supervision, Writing – Review & Editing; **Chantre-Astaiza A:** Methodology, Supervision, Writing – Review & Editing; **Ramirez-Gonzalez G:** Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2018 Rodríguez-Montúfar F *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

How to cite this article: Rodríguez-Montúfar F, Ordoñez-Buitron B, Duran D *et al.* **A structured process to create datasets with nutritional information [version 1; peer review: 2 not approved]** F1000Research 2018, 7:110 <https://doi.org/10.12688/f1000research.12979.1>

First published: 25 Jan 2018, 7:110 <https://doi.org/10.12688/f1000research.12979.1>

Introduction

In this day and age, there is an impressive amount of data traffic that is generated and shared over the internet. Researchers can utilize thousands of photos, hours of video footage, and consumer data to create datasets¹. Some datasets are used in research with a specific goal in mind, whereas other datasets are used to create data and store information for future investigations. Some datasets are freely published, while others are for restricted use.

There are several studies that use data to analyse taste preferences around online shopping², music^{3,4}, movies^{5,6} or social relations, for example ⁷. However, a study about people's preferences for food items in supermarkets in Colombia faces challenges due to the lack of datasets freely available on this topic. Additionally, various products that are present in some public datasets⁸ are not available in Colombia.

To address these gaps, this study describes the process of creating and describing a dataset that contains information on the food preferences and purchases of a group of people living in Colombia. An important aspect of the dataset is describing the sodium⁹ and sugar¹⁰ content of each food product and featuring and sorting out the nutritional information available in the Colombian market.

Methods

According to the STROBE guidelines, we have taken the following into consideration.

The purpose of the study is based on capturing the preferences of users in self-service stores. The study was carried out in the cities of Popayán and San Juan de Pasto, Colombia, across two months, where part of this period was used for participant recruitment.

A group of students, professionals and independent workers, all ≥ 18 years of age, accepted the invitation to participate in this research voluntarily, providing a signed agreement where they

accepted to sharing their information as long as their identities would be protected and remained anonymous.

All data were analyzed and stored in text files available in the “Variables and Data sources” section. In that section, the structure and components are explained in more detail.

The study is exploratory and the aim is to obtain a dataset for future work. The general structure followed the principles outlined by Robert K. Yin¹¹. **Table 1** presents a summary of these elements.

Data collection

Figure 1 illustrates the process of data acquisition, carried out using two methods. The first method involved collecting preferences using a survey, and the second method involved the acquisition of purchase records with invoices. All purchases were made in self-service stores, focused particularly on food self-service stores. Data collection was implemented over a one-month period when participants were actively involved in the data collection process.

User preferences

Data collection as described above was carried out through a survey, where people chose products based on their preferences. For this task, the Google Forms web tool was used, in which a series of questions were designed and classified into twelve sections. Participants were informed of the academic purpose of the survey, and the basic demographic data of each participant was registered. They identified their preferences out of the 708 food items presented in the survey. All items were classified into ten categories, created from observation of local self-service stores.

Table 2 shows the 10 categories the items were classified under. Classification of the items aimed to have participants interact in a more comfortable and conscious way with the questions, attempting to keep the process from becoming tedious.

Table 1. Summary of elements based on the Case Study Research by Robert K. Yin.

Phase	Summary
Plan	We want to answer the question: what are the items that people prefer when making purchases in self-service stores?
Design	The references in consumption of Items in a supermarket are selected as the unit of analysis. Type of simple case study Exploratory nature
Set Up	The structure proposed in Figure 2
Data Collection	Period of data collection: July-August, 2017
analysis	The data are available according to the structure proposed in Table 3 , Table 4 , Table 5 and Figure 1 , Figure 2
Release	Placing the data in the public domain by means of this article

Figure 1. Schematic of data collection.

Table 2. Names of categories used to classify products when listing user preference data.

SECTION	NAME OF THE SECTION
1	Groceries
2	Dairy products, sausages and chilled
3	Meat
4	Fruits and Vegetables
5	Fish and shellfish
6	Drinks
7	Liquors
8	Candies and snacks
9	Bakery
10	Frozen products

The survey was available for one month, and was available online. During the collection process, 215 people participated and shared their preferences and other demographic data.

User purchasing history

The purchase history refers to a list of products purchased by a person within a period of time in a self-service store. 65 participants provided all of their purchase receipts for four weeks, in particular for food products. At the end of this period, all the invoices of the 65 people who participated in the study were collected. R-Studio v1.0.143.

12 was then used to transcribe the products of interest, taking into account the number of submitted receipts, non-food products, and the number of times each user purchased each item.

Data treatment

The second part of Figure 1 illustrates how the information collected from the surveys was processed to construct the datasets. The process involved manually removing irrelevant information such as repeated surveys, inconsistent data, and non-focused responses in the user preferences section. For the participants' purchase receipts, some information was also manually removed, since some receipts contained purchases other than food products. The previously filtered information in both datasets was anonymized by assigning numerical codes to the users and the products to protect users' identities and classify all the products. All food items were classified based on their sodium and sugar content (based on WHO and FDA recommendations)^{9,10}. Figure 2 shows the final data structure after organizing the information¹³.

Figure 2. Schematic of data classification. Survey_items represents the preferences of the user and Purchase_items represents the purchases themselves, along with the characteristics of each product.

Table 3. User preferences dataset.

User Code	Item Code
2000000	1000002
2000000	1002003
2000000	1003006
2000000	1003006
2000001	1014001
2000001	1019005
.	.
.	.
2000213	1006015
2000214	1007000
2000214	1013001

Table 4. User purchasing history dataset.

User Code	Item Code	Rating
2000150	1052000	0.05
2000150	1013015	1
2000150	1056022	0.25
2000150	1056023	0.23
2000226	1034029	0.4
2000226	1002000	0.13
.	.	.
.	.	.
2000228	1056009	0.26
2000040	1000001	0.42
2000040	1059019	0.5

Data structure

There are two columns in **Table 3**. The first column (“User Code”) shows the code assigned to each user, and the second column registers the products selected as each user’s favorite. Each user has one or more products registered in the table, where the first four numbers represent the type of product, and the last three numbers refer to the specific brand for each product.

Similar to the previous table, **Table 4** presents the same two columns, and has an additional column, which shows the ranking, or the number of times a user has purchased that product divided by the number of shopping invoices for that user over the four week period.

Table 5 has six columns, with each row representing a different product characteristic. Each product has an item code, the section to which the product belongs, the category to which the product belongs, brand, sugar content per 100 g, and sodium content per serving (classified into four levels, where 1 is the lowest and 4 is the highest).

Ethical statement

Written informed consent was provided by all persons who volunteered in the research. Our study received approval in data management ethics according to the politics of the Telematics Engineering Group of the University of Cauca within the Electronics and Telecommunication Engineering Faculty. The proposed procedure is covered under approval number 8.4.2-90.14/274 of 2017. The study also complies with article 15 of the 1991 Colombian constitution on the right to privacy; and with the concepts of the Colombian Constitutional Court in Judgment No. T-414/92 of 1992 on the definition of data, computer freedom, and personal information.

Results

The numbers in the second half of **Figure 3** represent a scale for measuring sodium and sugar contents, based on the quantity

of sodium and sugar that each product contained according to the nutritional table. To better understand the graph, we note again that there are four levels that represent the sodium content, and four levels that represent the sugar content, which generates 16 possible combinations that are color coded differently. For instance, the green circle with the number 11 indicates that the sodium and sugar contents are very low, whilst the red circle with the number 44 indicates a product with very high sodium and sugar content. The pie chart illustrates the percentage of products in each sodium and sugar classification. There is a higher percentage of products with high sodium and low sugar contents.

Dataset 1. User preferences

<http://dx.doi.org/10.5256/f1000research.12979.d188373>

This file contains two columns (User_Code, Item_Code), the first column User_Code is the code assigned to each user and the second column Item_Code contains the encoded product that the user prefers.

Dataset 2. User purchasing

<http://dx.doi.org/10.5256/f1000research.12979.d188374>

This file contains three columns (User_Code, Item_Code, Rating), the first column User_Code is the code assigned to each user and the second column Item_Code contains the encoded product that the user prefers and Rating is the value obtained from dividing the number of total product invoices by the number of times the user purchased a product.

Dataset 3. Product characteristics

<http://dx.doi.org/10.5256/f1000research.12979.d188375>

This file contains six columns (Item_Code, Category, Section, Code_Brand, Sugar_Level, Sodium_Level). Item_Code is the code assigned to each item and the other columns represent how they have been classified and coded according to their characteristics.

Table 5. Product characteristics.

Item Code	Categ.	Sect.	Brand	Sug. level	Sodi. level
1058005	1	1100	2185	3	4
1052017	2	1204	2098	3	3
1039002	3	1301	2243	4	2
1045005	4	1401	2277	1	1
1018008	5	1504	2348	1	4
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
1032002	6	1604	2410	1	3
1037002	7	1700	2420	1	2
1060012	8	1800	2435	1	1
1034023	9	1900	2467	2	1

Figure 3. Levels of sodium and sugars in food and drink products.

Dataset 4. Products

<http://dx.doi.org/10.5256/f1000research.12979.d188376>

This file contains three columns (No, Item_Code, Product*), where Item_Code represents the code assigned to each product and Product represents the product type without specifying the brand.

Dataset 7. Sections

<http://dx.doi.org/10.5256/f1000research.12979.d188379>

This file contains three columns (No, Code_Section, Section). Code_Section represents the code assigned to each section and Section represents the section in which a product can be found.

Dataset 5. Brands

<http://dx.doi.org/10.5256/f1000research.12979.d188377>

This file contains three columns (No, Code_Band, Brand), Code_Band represents the code assigned to each brand and Brand represents the brand of each product.

Dataset 8. Sugar

<http://dx.doi.org/10.5256/f1000research.12979.d188380>

This file contains three columns (No, Sugar_Level, Code_Sugar_Level). Sugar_Level classifies the products by sugar content and Code_Sugar_Level represents the code assigned to each level.

Dataset 6. Categories

<http://dx.doi.org/10.5256/f1000research.12979.d188378>

This file contains three columns (No, Code_Category, Category), Code_Category represents the code assigned to each category and Category represents the assigned to the product.

Dataset 9. Sodium

<http://dx.doi.org/10.5256/f1000research.12979.d188381>

This file contains three columns (No, Sodium_Level, Code_Sodium_Level). Sodium_Level classifies the products by sodium content and Code_Sodium_Level represents the code assigned to each level.

Conclusions

This work was carried out to construct a valid dataset with food items available in Colombia. Future academic studies can perform statistical analysis using the data collected. Using the information from the nutritional labels of food items, we classified products using aspects like sodium and sugar content, following WHO and FDA recommendations to inform us whether the products contain levels above or below the recommended levels.

Data availability

Dataset 1: User preferences. This file contains two columns (User_Code, Item_Code), the first column User_Code is the code assigned to each user and the second column Item_Code contains the encoded product that the user prefers. DOI, [10.5256/f1000research.12979.d188373](https://doi.org/10.5256/f1000research.12979.d188373)¹⁴.

Dataset 2: User purchasing. This file contains three columns (User_Code, Item_Code, Rating), the first column User_Code is the code assigned to each user and the second column Item_Code contains the encoded product that the user prefers and Rating is the value obtained from dividing the number of total product invoices by the number of times the user purchased a product. DOI, [10.5256/f1000research.12979.d188374](https://doi.org/10.5256/f1000research.12979.d188374)¹⁵.

Dataset 3: Product characteristics. This file contains six columns (Item_Code, Category, Section, Code_Brand, Sugar_Level, Sodium_Level). Item_Code is the code assigned to each item and the other columns represent how they have been classified and coded according to their characteristics. DOI, [10.5256/f1000research.12979.d188375](https://doi.org/10.5256/f1000research.12979.d188375)¹⁶.

Dataset 4: Products. This file contains three columns (No, Item_Code, Product"), where Item_Code represents the code assigned to each product and Product represents the product type without specifying the brand. DOI, [10.5256/f1000research.12979.d188376](https://doi.org/10.5256/f1000research.12979.d188376)¹⁷.

Dataset 5: Brands. This file contains three columns (No, Code_Brand, Brand), Code_Brand represents the code assigned to each brand and Brand represents the brand of each product. DOI, [10.5256/f1000research.12979.d188377](https://doi.org/10.5256/f1000research.12979.d188377)¹⁸.

Dataset 6: Categories. This file contains three columns (No, Code_Category, Category), Code_Category represents the code assigned to each category and Category represents the assigned to the product. DOI, [10.5256/f1000research.12979.d188378](https://doi.org/10.5256/f1000research.12979.d188378)¹⁹.

Dataset 7: Sections. This file contains three columns (No, Code_Section, Section). Code_Section represents the code assigned to each section and Section represents the section in which a product can be found. DOI, [10.5256/f1000research.12979.d188379](https://doi.org/10.5256/f1000research.12979.d188379)²⁰.

Dataset 8: Sugar. This file contains three columns (No, Sugar_Level, Code_Sugar_Level). Sugar_Level classifies the products by sugar content and Code_Sugar_Level represents the code assigned to each level. DOI, [10.5256/f1000research.12979.d188380](https://doi.org/10.5256/f1000research.12979.d188380)²¹.

Dataset 9: Sodium. This file contains three columns (No, Sodium_Level, Code_Sodium_Level). Sodium_Level classifies the products by sodium content and Code_Sodium_Level represents the code assigned to each level. DOI, [10.5256/f1000research.12979.d188381](https://doi.org/10.5256/f1000research.12979.d188381)²².

Competing interests

No competing interests were disclosed.

Grant information

The author(s) declared that no grants were involved in supporting this work.

References

1. IBM.com: **What is a data set?**
[Reference Source](#)
2. Martinez-Pabon F, Ospina-Quintero JC, Ramirez-Gonzalez G, *et al.*: **Recommending ads from trustworthy relationships in pervasive environments.** *Mobile Information Systems*. 2016; **2016**: 18.
[Publisher Full Text](#)
3. Bertin-Mahieux T, Ellis DP, Whitman B, *et al.*: **The million song dataset.** In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*. 2011.
[Publisher Full Text](#)
4. Defferrard M, Benzi K, Vandergheynst P, *et al.*: **FMA: A Dataset For Music Analysis.** *ArXiv e-prints*. 2016.
[Reference Source](#)
5. Maxwell Harper F, Konstan JP: **The movielens datasets: History and context.** *Acm transactions on interactive intelligent systems (tiis)*. 2015; **5**(4): 19.
[Publisher Full Text](#)
6. Leka O: **IMDB movies dataset.** 2016.
[Reference Source](#)
7. Martinez-Pabon F, Caicedo-Guerrero J, Ibarra-Samboni JJ, *et al.*: **Smart TV-Smartphone Multiscreen Interactive Middleware for Public Displays.** *ScientificWorldJournal*. 2015; **2015**: 14, 534949.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. Pennacchioli D, Coscia M, Rinzivillo S, *et al.*: **Explaining the product range effect in purchase data.** 2013.
[Publisher Full Text](#)
9. Fda.gov: **Sodium in Your Diet: Use the Nutrition Facts Label and Reduce Your Intake.** Accessed 06-10-2017.
[Reference Source](#)
10. World Health Organization: **Sugars intake for adults and children.** 2015; **2015**: 59.
[Reference Source](#)
11. Yin RK: **Case Study Research.** 2002.
12. RStudio Team: **RStudio: Integrated Development Environment for R.** RStudio, Inc., Boston, MA, 2015.
13. TechTarget IncRouse M: **Class diagram.** 2017.
[Reference Source](#)
14. Rodriguez-Montufar F, Ordoñez-Buitron B, Duran D, *et al.*: **Dataset 1 in: A structured process to create datasets with nutritional information.** *F1000Research*. 2017.
[Data Source](#)
15. Rodriguez-Montufar F, Ordoñez-Buitron B, Duran D, *et al.*: **Dataset 2 in:**

- A structured process to create datasets with nutritional information.**
F1000Research. 2017.
[Data Source](#)
16. Rodriguez-Montufar F, Ordoñez-Buitron B, Duran D, *et al.*: **Dataset 3 in: A structured process to create datasets with nutritional information.**
F1000Research. 2017.
[Data Source](#)
17. Rodriguez-Montufar F, Ordoñez-Buitron B, Duran D, *et al.*: **Dataset 4 in: A structured process to create datasets with nutritional information.**
F1000Research. 2017.
[Data Source](#)
18. Rodriguez-Montufar F, Ordoñez-Buitron B, Duran D, *et al.*: **Dataset 5 in: A structured process to create datasets with nutritional information.**
F1000Research. 2017.
[Data Source](#)
19. Rodriguez-Montufar F, Ordoñez-Buitron B, Duran D, *et al.*: **Dataset 6 in: A structured process to create datasets with nutritional information.**
F1000Research. 2017.
[Data Source](#)
20. Rodriguez-Montufar F, Ordoñez-Buitron B, Duran D, *et al.*: **Dataset 7 in: A structured process to create datasets with nutritional information.**
F1000Research. 2017.
[Data Source](#)
21. Rodriguez-Montufar F, Ordoñez-Buitron B, Duran D, *et al.*: **Dataset 8 in: A structured process to create datasets with nutritional information.**
F1000Research. 2017.
[Data Source](#)
22. Rodriguez-Montufar F, Ordoñez-Buitron B, Duran D, *et al.*: **Dataset 9 in: A structured process to create datasets with nutritional information.**
F1000Research. 2017.
[Data Source](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 17 April 2018

<https://doi.org/10.5256/f1000research.14074.r31839>

© 2018 Kovalskys I. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Irina Kovalskys

International Life Sciences Institute (ILSI), Autonomous City of Buenos Aires, Buenos Aires, Argentina

The authors had a good idea and implemented accordingly. Despite of that, scientific information is weak and reproducibility is not ensured with the information provided

1. The goal of the study is not enough clear: Two phrases are used to describe the purpose:

a. "This study describes the process of creating and describing a dataset that contains information on the food preferences and purchases of a group of people living in Colombia. An important aspect of the dataset is describing the sodium and sugar content of each food product and featuring and sorting out the nutritional information available in the Colombian market.

b. The purpose of the study is based on capturing the preferences of users in self-service stores.

Which is the exact main goal? Describing the process of the data set or communicating the results obtained?

2. Methods

a. This publication should be improved by giving much more detail about sample.

Sample calculation and or sample size adequate to conclusions is missing.

Purchasing habits is directly related to socioeconomic status, culture and habits. It is not possible to know the population represented in this sample. (characteristics of the population are not detailed)

b. The authors say "All food items were classified based on their sodium and sugar content (based on WHO and FDA recommendations)". Food groups and food classification should be described in detail to better understand Figure 3

3. Conclusions

I suggest to first clarify the aim of the study to obtain the main conclusions. It would be expected to summarize the main outcomes in this section

I recommend the authors to write two scientific communications. One considering the process only and another one focusing on results

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

No

If applicable, is the statistical analysis and its interpretation appropriate?

Not applicable

Are all the source data underlying the results available to ensure full reproducibility?

No

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: NCD, obesity, nutrition, population studies

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 03 April 2018

<https://doi.org/10.5256/f1000research.14074.r31841>

© 2018 Gómez Salas G et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Georgina Gómez Salas

Departamento de Bioquímica, Escuela de Medicina, Universidad de Costa Rica, San José, Costa Rica

Yadira Cortes Sanabria

Pontificia Universidad Javeriana, Bogotá, Colombia

Even though the information obtained in this study could be used to further research, we consider that the simple attempt to provide a data base is not enough for publication. According to methodology, they should explain in detail about the sample, is socio-economic status a possible bias? How they were selected?

The paper lacks a discussion section to compare their results with other Colombian food data bases, or whether or not this can be extrapolated to other regions of the country.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Not applicable

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

No

Competing Interests: No competing interests were disclosed.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research