Journal of Food Composition and Analysis 22S (2009) S23–S31

Contents lists available at ScienceDirect

Journal of Food Composition and Analysis

journal homepage: www.elsevier.com/locate/jfca

Original Article

Classification of fruits and vegetables

Jean A.T. Pennington, Rachel A. Fisher *

Division of Nutrition Research Coordination, National Institutes of Health, Bethesda, MD 20892-5461, United States

ARTICLE INFO

Article history:
Received 18 June 2008
Received in revised form 12 November 2008
Accepted 18 November 2008

Keywords:
Antioxidants
Classification
Dietary assessment
Dietary guidance
Food composition
Fruits
Phytonutrients
TAC
Total antioxidant capacity
Vegetables

ABSTRACT

Classifications for fruits and vegetables are most helpful for dietary assessment and guidance if they are based on the composition of these foods. This work determined whether levels of food components in fruits and vegetables correlated with classification criteria based on botanic family, color, part of plant, and total antioxidant capacity (TAC). A database of 104 commonly consumed fruits and vegetables was created that contained food components known to be provided primarily by these foods. A mathematical clustering algorithm was used to group the foods into homogeneous clusters based on food component levels and the classification criteria. Most useful in categorizing were the botanic families rose, rue (citrus), amaryllis, goosefoot, and legume; color groupings blue/black, dark green/green, orange/peach, and red/purple; and plant parts fruit-berry, seeds or pods, and leaves. Groupings based on TAC levels did not match well with the identified clusters. Clusters were often best defined by a combination of classification variables such as color and part of plant. Results suggest that the groupings dark green leafy vegetables; cabbage family vegetables; lettuces; allium family bulbs; legumes; deep orange/yellow fruits, roots, and tubers; citrus family fruits; tomatoes and other red vegetables and fruits; and red/purple/blue berries are predictive for food components provided by fruits and vegetables.

Published by Elsevier Inc.

1. Introduction

Consumption of fruits and vegetables is important for human health because these foods are primary sources of some essential nutrients and contain phytochemicals that may lower risk of chronic disease (DGAC, 2005). Because the many fruits and vegetables available to the United States (US) population vary in composition for both nutrients and phytochemicals, classification of fruits and vegetables is important to researchers who attempt to assess relationships among diet, health, and disease. For example, researchers who develop food frequency questionnaires often need to gauge their fruit and vegetable questions to assess intakes of specific food components.

Classification of fruits and vegetables is also needed for dietary guidance materials to help people select appropriate types of these foods to meet their nutrient and health needs. Many countries have food guides with graphic depictions of the food groups and subgroups along with recommendations for consumption (Painter et al., 2002). The fruit and vegetable groups and subgroups vary from country to country because the focus of food guides is not only on the important components in fruits and vegetables but also on which fruits and vegetables are commonly available to and consumed by population groups.

Food guides created by the US Department of Agriculture (USDA) have included the Basic 7 Food Groups, the Basic 4 Food Groups, the Food Guide Pyramid, and, most recently, MyPyramid (CNPP, 2008). In the Basic 7 and Basic 4 food group guides there were recommendations to consume dark green leafy and deep-orange fruits and vegetables several times a week (for beta-carotene as precursor to vitamin A) and citrus fruits daily (for vitamin C). In the mid-1980s, a food guide from the National Cancer Institute (NCI) encouraged consumption of cruciferous vegetables (i.e. cabbage family vegetables) several times a week for their role in cancer prevention (NCI, 1986). Another NCI campaign called "5 A Day" encouraged classification of fruits and vegetables by color, but did not provide scientific background information that directly associated color to specific concentrations of food components. (The 5 A Day program was transferred to the Centers for Disease Control and Prevention and is now called "Fruits & Veggies-More Matters" (CDC, 2008). It no longer emphasizes the colors of fruits and vegetables, but is consistent with the MyPyramid (CNPP, 2008) recommendations.)

MyPyramid, which was issued in 2005, places all fruits together in one group and has five vegetable subgroups (Tables 1 and 2). The fruit group and vegetable subgroup consumption recommendations were based on the composition and national consumption patterns for food clusters consisting of 12 fruits and 36 different vegetables (Britten et al., 2006; Marcoe et al., 2006). The food components included in the development of the consumption recommendations included macronutrients, nine vitamins, and eight minerals.

* Corresponding author. Tel.: +1 301 594 7722; fax: +1 301 480 3768.
E-mail address: fisherrachel@mail.nih.gov (R.A. Fisher).

0889-1575/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.jfca.2008.11.012

Table 1
MyPyramid fruits and vegetables.

Group or subgroup	MyPyramid website examples (CNPP, 2008)[a]	MyPyramid composites (Marcoe et al., 2006)
Fruits	Apple, apricot, avocado, banana, berries (strawberries, blueberries, raspberries, cherries), grapefruit, grapes, kiwifruit, lemon, lie, mango, melon (cantaloupe, honeydew, watermelon), nectarine, orange, peach, pear, papaya, pineapple, plum, prune, raisins, and tangerine	12 fruits: banana, apple, watermelon, strawberries, grapes, cantaloupe, orange, peach, raisins, pear, plum, and grapefruit
Dark green vegetables	Boy choy, broccoli, collard greens, dark green leafy lettuce, kale, mesclun, mustard greens, romaine, spinach, turnip greens, and watercress	7 vegetables: raw and cooked broccoli, romaine, raw and cooked spinach, cooked collard greens, cooked mustard greens, cooked kale, and cooked turnip greens
Orange vegetables	Acorn squash, butternut squash, carrots, hubbard squash, pumpkin, and sweet potatoes	4 vegetables: raw and cooked carrots, cooked sweet potato, cooked winter squash, and cooked pumpkin
Dry beans and peas	Black beans, black-eyed peas, garbanzo beans (chickpeas), kidney beans, lentils, lima beans (mature), navy beans, pinto beans, soy beans, split peas, tofu (bean curd made from soybeans), and white beans	10 vegetables: pinto beans, white beans, soybeans, kidney beans, black beans, lentils, chickpeas, lima beans, cowpeas, and split peas
Starchy vegetables	Corn, green peas, lima beans (green), and potatoes	3 vegetables: boiled and baked potato, corn, and green peas
Other vegetables	Artichokes, asparagus, bean sprouts, beets, Brussels sprouts, cabbage, cauliflower, celery, cucumber, eggplant, green beans, green or red peppers, iceberg lettuce, mushrooms, okra, onions, parsnips, tomato, tomato juice, vegetable juice, turnip, wax beans, and zucchini	12 vegetables: raw and cooked tomato and tomato juice, lettuce, raw cucumber, raw pepper, raw and cooked onion, raw and cooked celery, raw and cooked cabbage, cooked green beans, mushroom, cooked bean sprouts, cooked summer squash, and cooked cauliflower

[a] Vegetables may be raw or cooked; fresh, frozen, canned, or dried/dehydrated; and may be whole, cut-up, or mashed. Fruit may be fresh, canned, frozen, or dried, and may be whole, cut-up, or pureed. MyPyramid provides recommendations for daily intakes for fruits and vegetables and weekly intakes from the five vegetable subgroups. Underlined foods are those included in the composites used to develop the food group recommendations.

Other systems for classifying fruits and vegetables that appear to be related to food composition include those based on botanic families, colors, and edible parts of plants (Pennington, 2003). Another potentially important classification might be total antioxidant capacity (TAC) because the protection that fruits and vegetables offer against cancer, cardiovascular diseases, and other diseases has been attributed to various antioxidants they contain (Wang et al., 1996; Wu et al., 2004). TAC is thought to reflect concentrations of ascorbic acid, alpha-tocopherol, beta-carotene, various flavonoids, glutathione, and other (mainly phenolic) compounds present in foods.

The goal of this work was to group fruits and vegetables into homogeneous clusters based on their food component profiles and investigate potential classification variables for these foods that relate specifically to food composition (irrespective of food consumption). Such classifications could help researchers develop the fruit and vegetable sections of their food frequency questionnaires, teach students about food composition, and assist dietitians in providing dietary guidance to patients and clients. Previous studies suggest that mathematical clustering algorithm methods can be used to overcome the challenge of dealing with multiple food components simultaneously to objectively classify foods (Akbay et al., 2000; Windham et al., 1985).

Table 2
MyPyramid fruit and vegetable clusters (Marcoe et al., 2006).

Fruit or vegetable	Foods represented by fruit or vegetable
Banana	Plantain
Strawberries	Kiwifruit, blueberries, cranberries, raspberries, blackberries
Grapes	Cherries, rhubarb
Cantaloupe	Honeydew melon, casaba melon
Orange	Tangerine
Peach	Mango, papaya, apricot, guava, avocado
Raisins	Date, fig
Plum	Prune
Romaine	Endive, chicory, escarole, parsley
Kale	Chard, parsley, dandelion
Soybeans	Tofu, soy-based meal replacements, soy beverages
Boiled potato	Fried potatoes, chips
Cooked tomato	Tomato sauce
Lettuce	Iceberg and butterhead lettuce
Pepper	Green, red, and chili peppers; olives
Onion	Mature and green onions, leeks, chives, garlic
Cabbage	Red and green cabbage
Green beans	Snow peas, asparagus, okra, artichokes
Summer squash	Zucchini

2. Methods and materials

2.1. Fruit and vegetable definitions

Although the botanic term "fruit" refers to the seeds and surrounding tissues of a plant, the foods that are commonly referred to as "fruits" for culinary purposes are pulpy seeded tissues that have a sweet (oranges, apples, pears, blueberries) or tart (lemons, limes, cranberries) taste. In the US diet, fruits are often consumed as snacks, desserts, or a sweet side dish to a meal.

By culinary definition, "vegetables" are edible plant parts including stems and stalks (celery), roots (carrots), tubers (potatoes), bulbs (onions), leaves (spinach, lettuce), flowers (artichokes), some fruits (cucumbers, pumpkin, tomatoes), and seeds (beans, peas). Vegetables are less sweet or tart than "fruits" and are usually consumed as salads, cooked side dishes, and savory appetizers. Using this definition, avocado was grouped with vegetables for this work. Mushrooms (fungi) and sweet corn (a cereal grain) are included as vegetables here because they are commonly used as vegetables in US diets. The term "legumes" is used here to mean beans and peas; however, peanuts (a type of legume used as nuts) are not included. Herbs (e.g. coriander, basil) and vegetables used primarily as spices (e.g. chives) are not included here because they are generally used in small amounts and were assumed not to contribute greatly to food component intake.

2.2. Identifying fruits and vegetables for the database

Table 3 provides alphabetical lists for the 37 fruits and 67 vegetables selected for the database. The goals for the database were to include only commonly consumed fruits and vegetables, to have only one version (i.e. one listing) of each fruit and vegetable, to have a complete food component profile for each food (i.e. no

Table 3
Fruits and vegetables in the database.

Fruits	
1	Apple, raw, with skin
2	Apricot, raw
3	Asian pear, raw
4	Banana, raw
5	Blackberries, raw
6	Blueberries, raw
7	Boysenberries, frozen, unsweetened
8	Cantaloupe, raw
9	Casaba melon, raw
10	Cherries, sweet, raw
11	Clementine, raw
12	Cranberries, raw
13	Date, raw
14	Fig, raw
15	Grapefruit, pink/red, raw
16	Grapefruit, white, raw
17	Grapes, red, raw
18	Guava, raw
19	Honeydew melon, raw
20	Kiwifruit, raw
21	Kumquat, raw
22	Lemon, raw
23	Lime, raw
24	Mango, raw
25	Nectarine, raw
26	Orange, raw
27	Papaya, raw
28	Peach, raw
29	Pear, raw
30	Pineapple, raw
31	Plum, raw
32	Pomegranate, raw
33	Raisins, seedless
34	Raspberries, raw
35	Strawberries, raw
36	Tangerine (mandarin orange), raw
37	Watermelon, raw
Vegetables	
1	Artichoke, raw
2	Asparagus, raw
3	Avocado, raw
4	Beet greens, raw
5	Beet, raw
6	Blackeye peas (cowpeas), immature seeds, raw
7	Broccoli raab (rapini), raw
8	Broccoli, raw
9	Brussels sprouts, raw
10	Butternut squash, raw
11	Cabbage, green, raw
12	Cabbage, red, raw
13	Carrot, raw
14	Cauliflower, raw
15	Celery, raw
16	Chinese broccoli, boiled
17	Chinese cabbage (pak choi/bok choy), raw
18	Collards, raw
19	Corn, sweet, yellow, raw
20	Crookneck/straightneck squash, raw
21	Cucumber, raw
22	Eggplant, raw
23	Endive, raw
24	Garlic, raw
25	Hubbard squash, raw
26	Jerusalem artichoke (sunchoke), raw
27	Jicama (yambean), raw
28	Kale, raw
29	Kidney beans, mature seeds, red, raw
30	Leek, raw
31	Lentils, raw
32	Lettuce, butterhead, raw
33	Lettuce, iceberg, raw
34	Lettuce, green leaf, raw
35	Lettuce, red leaf, raw
36	Lima beans, immature seeds, raw
37	Mung beans, mature seeds, raw
38	Mushrooms, common, raw
39	Mustard greens, raw
40	Navy beans, mature seeds, raw
41	Okra, raw
42	Onion, raw
43	Parsley, raw
44	Parsnip, raw
45	Peas, green, raw
46	Pepper, sweet, green, raw
47	Pepper, sweet, red, raw
48	Pigeon peas, immature seeds, raw
49	Pinto beans, mature seeds, raw
50	Potato with peel, raw
51	Pumpkin, raw
52	Radish, raw
53	Rhubard, raw
54	Romaine (cos lettuce), raw
55	Rutabaga, raw
56	Scallion (spring onion), raw
57	Snap beans, green, raw
58	Snowpeas, raw
59	Soybeans, mature seeds, raw
60	Spinach, raw
61	Sweet potato with peel, raw
62	Swiss chard, raw
63	Tomato, red, raw
64	Turnip, raw
65	Turnip greens, raw
66	Watercress, raw
67	Zucchini, raw

missing values), and to have complete information on the variables of interest (botanic family, color, part of plant, and TAC). The USDA Nutrient Database for Standard Reference, Release 20 (SR20) (ARS, 2008) was the primary data source for food component data. Some fruits and vegetables originally selected for the database were omitted because of too many missing values in SR20.

The foods in the database are in the raw form with three exceptions. Chinese broccoli is cooked (raw data were not available in SR20), boysenberries are frozen (raw data were not available in SR20), and raisins are dried. The raw forms were otherwise used because there are generally more data for raw fruits and vegetables than for cooked or processed versions, and also because it was desirable to stay close to inherent composition and not bring in the variables of food component losses or gains during processing and the potential effects of compositional changes on the correlations.

The data are provided per 100 g of food, rather than serving size, so that the correlations are not distorted by differences in weights for one-half or one cup serving portions or by different sizes and weights of whole fruits and vegetables.

2.3. Food components included and sources of data

The database includes food components that are known to be provided primarily by fruits and vegetables in US diets, i.e. dietary fiber, calcium, iron, magnesium, potassium, zinc, copper, manganese, vitamin C, vitamin B6, folate, vitamin A, vitamin K, phystosterol, alpha- and beta-carotene, lycopene, lutein + zeaxanthin, anthocyanidin, flavan-3-ols, flavonones, flavones, and flavonols. As noted, most of the data are from SR20. Missing values (primarily for phytosterol) were obtained from other sources (Duester, 2001; Norman et al., 1999; Ruiz et al., 2005; Ryan et al., 2007; Schutz et al., 2006; Souci et al., 2000; Weihrauch and Chatra, 1994) or imputed using values for similar foods.

2.4. Variables—Botanic family, color, part of plant, and TAC

2.4.1. Botanic family

Botanic classification of plants is based on the physiologic characteristics of plant development, organization, and structure.

Table 4
Fruits classified according to botanic family.

	Botanic family[a]	Fruit members	#
1	(Actinidiaceae)	Kiwifruit	1
2	Cashew (Anacardiacea)	Mango	1
3	Pineapple (Bromeliaceae)	Pineapple	1
4	Carica (Caricaceae)	Papaya	1
5[b]	Gourd (Cucurbitaceae)	Cantaloupe, casaba melon, honeydew melon, watermelon	4
6	Health (Ericaceae)	Blueberries, cranberries	2
7	Mulberry (Moraceae)	Fig	1
8	Banana (Musaceae)	Banana	1
9	(Myrtaceae)	Guava	1
10	Palm (Palmaceae)	Date	1
11	Pomegranate (Punicaceae)	Pomegranate	1
12[b]	Rose (Rosaceae)	Apple, apricot, Asian pear, blackberries, boysenberries, cherries, nectarine, peach, pear, plum, raspberries, strawberries	12
13[b]	Rue (Rutaceae)	Clementine, grapefruit (pink/red, white), kumquat, lemon, lime, orange, tangerine	8
14	Grape (Vitaceae)	Grapes/raisins (dried grapes)	2
			37

[a] The common botanic family name is listed first, if known, and the Latin botanic family name is in parentheses.
[b] Possible classification variables.

The family is the level before plants are given genus and species names. Botanic classification is useful for biologists to establish plant origins and relationships and to help identify plants among different cultures and languages. It is also useful for horticulturists because plants within a family may have similar climatic requirements, economic uses, and disease and insect controls. The usefulness of botanic classification to serve the needs of nutritionists is more complex because foods derived from the same botanic family may or may not contain similar levels of food components.

Table 4 lists the 14 botanic families for the 37 fruits in the database. Pennington (2003) was used as the source of botanic family information. Ten of the families contained only one food and could, therefore, not be used as predictors of composition. The gourd, health, rose, and rue families had at least two different fruit members.

Table 5 lists the 15 botanic families for the 67 vegetables. Seven of the families had only one vegetable member. The only botanic family that contained both fruits and vegetables was the gourd family. Families that might contain vegetables with somewhat similar food component profiles include amaryllis, sunflower, goosefoot, sunflower, cabbage, and legume. Foods within the gourd, nightshade, and carrot families do not contain similar food component profiles. The gourd family includes cantaloupe, watermelon, pumpkin, winter squashes, honeydew melon, summer squashes; the nightshade family includes chili peppers, sweet peppers, tomatoes, eggplant, and white potatoes; and the carrot family includes celery, celeriac, and parsnip.

Another issue that makes use of botanic families somewhat difficult for classifying fruits and vegetables is that multiple parts of some plants are eaten as separate vegetables and have very different food components (e.g. beet roots and greens, turnip roots and greens, broccoli stems and flowers, chive bulbs and green tops). Botanic classification applies to the entire plant and is not specific for different parts of the plant that are typically consumed.

2.4.2. Color

Fruits and vegetables come in a wide variety of colors and shades. The colors reflect the pigments present in the tissues, and some food components are known to be pigmented, e.g. beta-carotene is deep orange, and anthocyanidin is red. The green

Table 5
Vegetables classified according to botanic family.

	Botanic family[a]	Vegetable members	#
1[b]	Amaryllis (Amaryllidaceae)	Garlic, leek, onion, scallion	4
2[b]	Sunflower (Asteraceae/Compositae)	Artichoke, endive, Jerusalem artichoke, lettuce (butterhead, iceberg, green leaf, red leaf), romaine	8
3[b]	Goosefoot (Chenopodiaceae)	Beet and beet greens, spinach, Swiss chard	4
4	Morning glory (Convolvulaceae)	Sweet potato	1
5[b]	Cabbage (Cruciferae/Brassica)	Broccoli, broccoli raab, Brussels sprouts, cabbage (green, red), cauliflower, Chinese broccoli, Chinese cabbage, collards, kale, mustard greens, radish, rutabaga, turnip and turnip greens, watercress	16
6[b]	Gourd (Cucurbitaceae)	Butternut squash, crookneck/straightneck squash, cucumber, hubbard squash, pumpkin, zucchini	6
7	Fungi	Mushrooms	1
8	(Gramineae/Poaceae)	Sweet corn	1
9	Laurel (Lauraceae)	Avocado	1
10[b]	Legume (Leguminosae)	Blackeye peas, green peas, jicama, kidney beans, lentils, lima beans, mung beans, navy beans, pigeon peas, pinto beans, green snap beans, snowpeas, soybeans	13
11	Asparagus (Liliaceae)	Asparagus	1
12	Mallow/cotton (Malvaceae)	Okra	1
13	(Portulacacea)	Rhubarb	1
14[b]	Nightshare (Solanaceae)	Eggplant, sweet green pepper, sweet red pepper, potato, tomato,	5
15[b]	Carrot/parsley (Umbelliferae)	Carrot, celery, parsley, parsnip	4
			67

[a] The common botanic family name is listed first, if known, and the Latin botanic family name is in parentheses.
[b] Possible classification variables.

Table 6
Fruits and vegetables classified according to nine color groupings.

Color(s)	Fruits and Vegetables (#)	#
White/beige	Apple, Asian pear, banana, white grapefruit, pear (5) Artichoke, blackeye peas, cauliflower, crookneck/straightneck squash, cucumber, eggplant, endive, garlic, Jerusalem artichoke, jicama, leek, lentils, mushrooms, navy beans, onion, parsnip, pigeon peas, pinto beans, potato, radish, scallion, soybeans, turnip, zucchini (24)	29
Orange/peach	Apricot, cantaloupe, clementine, guava, kumquat, mango, nectarine, orange, papaya, peach, tangerine (11) Butternut squash, carrot, hubbard squash, pumpkin, sweet potato (5)	16
Yellow	Lemon, pineapple (2) Avocado, sweet corn, rutabaga (3)	5
Blue/black	Blackberries, blueberries, boysenberries, date, raisins (5)	5
Red/purple	Cherries, cranberries, grapes, plum, pomegranate, raspberries, strawberries, watermelon (8) Beet, red cabbage, kidney beans, red leaf lettuce, rhubarb, sweet red pepper, tomato (7)	15
Pink	Fig, pink/red grapefruit (2)	2
Light green	Casaba melon, honeydew melon, kiwifruit, lime (4) Cabbage, celery (2)	6
Green	Asparagus, broccoli, broccoli raab, Brussels sprouts, Chinese broccoli, Chinese cabbage, green peas, lettuce (iceberg, butterhead, green leaf), lima beans, mung beans, okra, parsley, romaine, snowpeas, sweet green pepper, green snap beans, watercress (19)	19
Dark green	Beet greens, collards, kale, mustard greens, spinach, Swiss chard, turnip greens (7)	7

color of plant foods is due to chlorophyll, which is not an essential or beneficial component for humans; however, chlorophyll masks the color of beta-carotene, so a green or deep green color is usually a predictor for beta-carotene.

To minimize the number of color groupings and thus enhance potential correlations, the primary color(s) for each fruit and vegetable was used. The 37 fruits and 67 vegetables were organized according to the nine color groups shown in Table 6. The color reflects the primary edible portion of each fruit and vegetable. The peel color was not reported if the primary edible portion was of a different color. For example, apples, zucchini, and eggplant are "white/beige" rather than the color of their peels. For one fruit and three vegetables, there are two listings based on color (e.g. pink/red and white grapefruit, green and red cabbage, green and red leaf lettuce, green and red sweet pepper). These are commonly consumed foods, and the colors are known to affect the levels of some food components.

2.4.3. Part of plant

Classification by edible part attempts to group fruits and vegetables by the part of the plant, bush, vine, or tree that is used as food. Similarity in composition among plant parts may be due to the functions of these tissues. Stem and stalk vegetables are usually high in dietary fiber, which serves to support the structure of the plant. Leaves, especially the dark green ones, tend to be the most metabolically active and most nutritious part of plants and are usually good sources of dietary fiber, folate, carotenoids, vitamin C, flavonoids, and minerals such as iron, zinc, calcium, and magnesium. Legumes are sources of protein, starch, isoflavones, vitamin B6, folacin, iron, and other minerals. Bulbs are noted for allicin, a food component related to cancer prevention (Lampe, 1999). Enlarged roots and tubers are storage organs for plants and some are high in starch content. Table 7 presents the six parts of plant groupings for vegetables (fruit; seeds or pods; stem, stalk, and/or flower; root or tuber; leaves; and bulb). One subgroup for the fruits was also identified, berries, in an attempt to see if this type of fruit might have some correlations different from other fruits.

2.4.4. TAC

TAC is a measure that shows the efficiency of antioxidant compounds in scavenging free radicals. Its value depends on the quality and quantity of the antioxidants contained in fruits and

Table 7
Fruits and vegetables classified according to part of plant.

Plant part	Fruits and vegetables	#
Fruit	Apple, apricot, Asian pear, banana, blackberries, blueberries, boysenberries, cantaloupe, casaba melon, cherries, Clementine, cranberries, date, fig, grapefruit (pink/red, white), grapes, guava, honeydew melon, kiwifruit, kumquat, lemon, lime, mango, nectarine, orange, papaya, peach, pear, pineapple, plum, pomegranate, raisins, raspberries, strawberries, tangerine, watermelon Avocado, butternut squash, crookneck/straightneck squash, cucumber, eggplant, hubbard squash, okra, sweet pepper (green, red), pumpkin, tomato, zucchini Berry subgroup: blackberries, blueberries, boysenberries, raspberries, strawberries	49
Seeds or Pods	Blackeye peas, corn, green peas, kidney beans, lentils, lima beans, mung beans, navy beans, pigeon peas, pinto beans, green snap beans, snowpeas, soybeans	13
Stem, stalk, and/or flower	Artichoke, asparagus, broccoli, broccoli raab, cauliflower, celery, Chinese broccoli, mushrooms[a], rhubarb	9
Root or tuber	Beet, carrot, Jerusalem artichoke, jicama, parsnip, potato, radish, rutabaga, sweet potato, turnip	10
Leaves	Beet greens, Brussels sprouts, cabbage (green, red), Chinese cabbage, collards, endive, kale, leek, lettuce (butterhead, iceberg, green leaf, red leaf), mustard greens, parsley, romaine, spinach, Swiss chard, turnip greens, watercress	20
Bulb	Garlic, onion, scallion	3

[a] Stem, stalk, and/or flower vegetable seemed the best fit for mushrooms, which are fungi and not plants.

Table 8
Fruits and vegetables classified according to six TAC[a] groupings.

TAC (μmol TE/100g)[a]	Fruits and vegetables (#)	#
>9000	Blueberries, cranberries (2) Artichoke, kidney beans, pinto beans, soybeans (4)	6
>6000, <9000	Guava, plum (2) Parsley (1)	3
>3000, <6000	Apple, blackberries, boysenberries, cherries, pomegranate, raisins, raspberries, strawberries (8) Asparagus, blackeye peas, broccoli raab, turnip greens (4)	12
>1000, <3000	Apricot, clementine, date, fig, pink grapefruit, grapes, lemon, lime, mango, orange, papaya, peach, Asian pear, pear, tangerine (15) Avocado, beet, beet greens, broccoli, cabbage (red, green), carrot, Chinese broccoli, Chinese cabbage, collards, eggplant, garlic, Jerusalem artichoke, jicama, kale, lettuce (butterhead, green leaf, red leaf), mung beans, mustard greens, navy beans, onion, parsnip, potato, rutabaga, scallion, spinach, Swiss chard, turnip (29)	44
>500, <1000	Banana, kiwi fruit, kumquat, nectarine, pineapple (5) Brussels sprouts, cauliflower, celery, corn, iceberg lettuce, lentils, mushrooms, green peas, sweet pepper (green, red), pigeon peas, radish, rhubarb, romaine, snowpeas, sweet potato (16)	21
<500	Cantaloupe, casaba melon, white grapefruit, honeydew melon, watermelon (5) Butternut squash, crookneck/straightneck squash, cucumber, hubbard squash, leek, iceberg lettuce, lima beans, okra, pumpkin, green snap beans, tomato, watercress, zucchini (13)	18

[a] TAC, Total antioxidant capacity expressed as micromoles of Trolox Equivalent (TE) per 100 g of food.

vegetables. Information about TAC for the 37 fruits and 67 vegetables was obtained from the literature (Cao et al., 1996; Guo et al., 2003; Miller et al., 2000; Soong and Barlow, 2004; Wang et al., 1996; Wu et al., 2004); missing values were imputed based on similar foods. TAC is expressed as micromoles of Trolox Equivalents per 100 g of fruit or vegetable. Table 8 groups the fruits and vegetables according to six TAC levels.

2.5. Statistical analysis

A mathematical clustering algorithm was used to group fruits and vegetables on the basis of similarity in the content of 23 food components known to be provided primarily by these foods in US diets. The empirically derived clusters were then compared to the four classification variables listed in Section 2.4 to determine if or how these variables could be used to classify fruit and vegetables groups. All computation was carried out using SAS (version 9, SAS Institute Inc., Cary, NC).

Data were analyzed using an agglomerative hierarchical clustering method. A benefit of this method is that clusters are not affected by the order in which the data are presented. Additionally, an initial guess at the number of clusters present in the data is not required, though a solution must still be selected from the nested hierarchy that results. Prior to analysis, values for all 23 food components were standardized using the range method. Because clustering is highly dependent on the selected units of measurement, standardization is usually recommended prior to the application of the clustering algorithm to control for different measurement units used for each data element. To determine the association between the individual sample units, the Euclidean distance similarity metric was chosen (SAS Institute Inc., 2005). Clusters were then joined using the linkage method of Ward's minimum variance (SAS Institute Inc., 2005).

Results were assessed through both graphical and statistical methods. A dendrogram (a tree of the sequence of cluster fusions) was generated to examine the feasible range of values for the number of clusters present in the data. The pseudo t2 statistic was also used to assist in the determination of the optimal number of clusters by measuring the separation between the two clusters most recently joined (SAS Institute Inc., 2005).

The derived clusters were profiled in several ways to determine the best descriptor for each group. First, matrices that depict the clusters against each class variable (botanic family, color, part of plant, and TAC) were created to determine which classification

Table 9
Fruit and vegetables according to clusters.

Cluster	Fruit and vegetables	#
1 (Other)	Pear, Asian pear, apricot, nectarine, peach, apple, plum, mango, pomegranate, banana, fig, cherries, grapes, raspberries, strawberries, casaba melon, honeydew melon, cantaloupe, rhubarb, pineapple, kiwi, papaya, raisins, date, crookneck squash, hubbard squash, zucchini, eggplant, turnip, jicama, beet, parsnip, potato, rutabaga, radish, corn, celery, cucumber, iceberg lettuce, butterhead lettuce, endive, snap beans, Jerusalem artichoke, artichoke, mushrooms, asparagus, snowpeas, scallion, leek, avocado, red pepper, green pepper, green cabbage, cauliflower, broccoli, Brussels sprouts, broccoli raab, Chinese broccoli, red cabbage	59
2 (Citrus)	Tangerine, clementine, lemon, lime, white grapefruit, orange, kumquat, pink grapefruit	8
3 (Dark berry)	Blackberries, boysenberries, blueberries	3
4 (Green leafy/orange vegetables)	Green leaf lettuce, red leaf lettuce, Romaine lettuce, Chinese cabbage, watercress, mustard greens, spinach, turnip greens, beet greens, Swiss chard, collards, pumpkin, butternut squash, sweet potato, carrots	15
5 (Peas/lentils/onion)	Mung beans (mature), green peas, pigeon peas (immature), blackeye peas (immature), lentils, lima beans (immature), okra, garlic, onion, cranberries	10
6 (Legumes)	Kidney beans (mature), pinto beans (mature), navy beans (mature), soybeans (mature)	4
7 (Lycopene/red)	Watermelon, guava, tomato	3
8 (Parsley/kale)	Kale, parsley	2

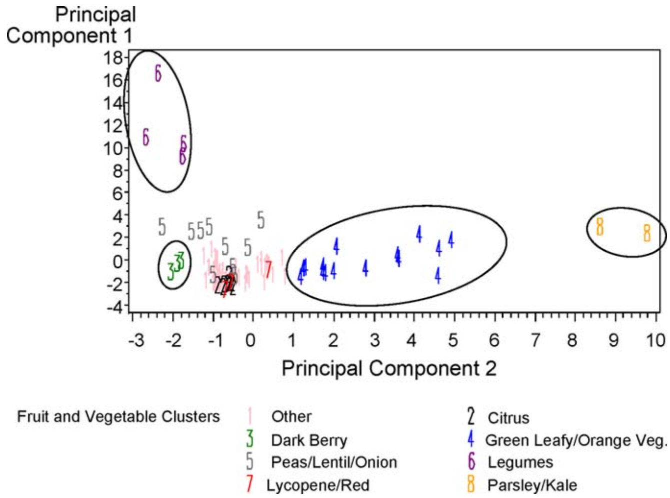

Fig. 1. Fruit and vegetable principal component plot.

variable(s) best represented the clusters. Second, differences between cluster means for each of the 23 input variables were compared with a multivariate analysis of variance (MANOVA) performed by PROC GLM (version 9, SAS Institute Inc., Cary, NC). Pair-wise comparisons were generated using multiple contrasts. To identify the input variables contributing to the differences, an analysis of variance (ANOVA) was performed with a Tukey's Studentized Range test ($P < 0.05$).

3. Results

Using the hierarchical clustering method with Ward's linkage, eight clusters were identified that could be used to classify fruits and vegetables. Cluster membership is displayed in Table 9. A principal component plot was created to give a sense of how many well-separated clusters existed in the data. As can be seen in Fig. 1, four circled clusters (clusters 3, 4, 6, and 8) are well separated when plotted against the two chosen principal components, which suggest that the food component profiles of these clusters are unique. There is overlap among the other four clusters (clusters 1, 2, 5, and 7) showing less distinction. Plots whose axes are formed by other principal components show better separation of the groups that appear overlapped in this particular plot.

3.1. Predictive value of botanic family, color, part of plant, and TAC

Matrices depicting the clusters against each class variable (botanic family, color, part of plant, and TAC) can be seen Tables 10–13. These matrices help to determine cluster labels. Any interpretable cluster solution should have a class label that can be used to define it. In this case, seven of the eight clusters could be defined with such a label. The more ambiguous cluster 1 or "other" category suggests that there are less discerning characteristics among these fruits and vegetables when compared to the other items in this database.

Cluster 2 contains 100% of the rue botanic family and contains no other members. Cluster 2 was accordingly labeled the "citrus" cluster (Table 10). Cluster 3 contains 50% of the fruit-berry category, of which all have a blue/black color. This cluster was thus labeled the "dark berry" cluster (Table 11). Sixty percent of the fruits and vegetables in the leafy group fell in Cluster 4 (Table 12), as did 80% of all the orange vegetables. It is, thus, labeled "green/dark green leafy vegetables/dark orange vegetables." Cluster 4 is significantly higher in beta-carotene, vitamin A, and vitamin K than all other clusters except Cluster 8. Cluster 5 contains 50% of all seeds/pods and 46% of the legumes. It also represents 50% of the amaryllis family, and is described as the "peas/lentils/onion" cluster. All members of Cluster 6 are part of the legume family, although it represents less than half of the legumes in the dataset. However, three out of the four members of Cluster 6 have a TAC in the highest category (>9000 μmol Trolox Equivalents/100 g; Table 13). The legumes that clustered in this group were all mature legumes, which have a lower water content and therefore higher food component concentrations than immature legumes. The legumes in Cluster 5 tended to be immature, so it is possible that Clusters 5 and 6 are separated only because of water content or some other characteristic of legume maturity. Cluster 7 is distinguished from the other clusters due to the high lycopene content of its members. All members of Cluster 7 were the fruiting part of the plant, and the color of the edible portion was red or peach. Cluster 8 is small and very similar to Cluster 4, though it is set apart by its members' significantly higher content of

Table 10
Botanic families by clusters[a].

Clusters	1 (Other)	2 (Citrus)	3 (Dark berry)	4 (Green leafy/orange vegetables)	5 (Peas/lentils/onion)	6 (Legumes)	7 (Lycopene/red)	8 (Parsley/kale)
Rose (Rosaceae)	10	0	2	0	0	0	0	0
Cabbage (Cruciferae)	10	0	0	5	0	0	0	1
Gourd (Cucurbitaceae)	7	0	0	2	0	0	1	0
Sunflower (Asteraceae/ Compositae)	5	0	0	3	0	0	0	0
Nightshade (Solanaceae)	4	0	0	0	0	0	1	0
Legume (Leguminosae)	3	0	0	0	6	4	0	0
Amaryllis (Ameryllidaceae)	2	0	0	0	2	0	0	0
Carrot/parsley (Umbelliferae)	2	0	0	1	0	0	0	1
Grape (Vitaceae)	2	0	0	0	0	0	0	0
Goosefoot (Chenopodiaceae)	1	0	0	3	0	0	0	0
Heath (Ericaceae)	0	0	1	0	1	0	0	0
Rutaceae	0	8	0	0	0	0	0	0

[a] The common botanic family name is listed first, if known, and the Latin botanic family name is given in parentheses. Botanic families represented by only one fruit or vegetable are not included in this table.

Table 11
Color by cluster.

Clusters	1 (Other)	2 (Citrus)	3 (Dark berry)	4 (Green leafy/orange vegetables)	5 (Peas/lentils/onion)	6 (Legumes)	7 (Lycopene/red)	8 (Parsley/kale)
White/beige	20	1	0	0	5	3	0	
Red/purple	10	0	0	1	1	1	2	
Green	10	0	0	4	4	0	0	1
Orange/peach	7	4	0	4	0	0	1	
Light green	5	1	0	0	0	0	0	
Yellow	4	1	0	0	0	0	0	
Blue/black	2	0	3	0	0	0	0	
Pink	1	1	0	0	0	0	0	
Dark green	0	0	0	6	0	0	0	1

Table 12
Part of plant by cluster.

Clusters	1 (Other)	2 (Citrus)	3 (Dark berry)	4 (Green leafy/orange vegetables)	5 (Peas/lentils/onion)	6 (Legumes)	7 (Lycopene/red)	8 (Parsley/kale)
Fruit	29	8	0	2	0	0	3	0
Stem, stalk and/ or flower	9	0	0	0	0	0	0	0
Root or tuber	8	0	0	2	0	0	0	0
Leaves	6	0	0	11	0	0	0	2
Seeds or pods	3	0	0	0	7	4	0	0
Bulb	2	0	0	0	2	0	0	0
Fruit-berry	2	0	3	0	1	0	0	0

Table 13
TAC[a] by cluster.

Clusters	1 (Other)	2 (Citrus)	3 (Dark berry)	4 (Green leafy/orange vegetables)	5 (Peas/lentils/onion)	6 (Legumes)	7 (Lycopene/red)	8 (Parsley/kale)
>9000	1	0	1	0	0	3	0	1
6000–<9000	1	0	0	0	0	0	1	0
3000–<6000	8	0	2	0	1	0	0	0
1000–<3000	25	6	0	10	3	1	0	1
500–<1000	15	1	0	2	3	0	0	0
<500	9	1	0	3	2	0	2	0

[a] TAC, Total Antioxidant Capacity expressed as micromoles of Trolox Equivalent (TE) per 100 g of food.

beta-carotene, vitamin A, lutein and zeaxanthin, vitamin C, and flavones. It is referred to as the "parsley/kale" cluster.

3.2. Suggested groupings

Because some lesser-known food components were not included in the database, they had no impact on cluster formation. To factor some of these "missing" food components into the final fruit and vegetable classification scheme, some small modifications were made to the empirical results of the cluster analysis to arrive at the following suggested 10 categories of fruits and vegetables.

- Dark green leafy vegetables
- Cabbage family vegetables
- Lettuces
- Allium family bulbs
- Legumes
- Deep orange/yellow fruits, roots, and tubers
- Citrus family fruits
- Tomatoes and other red vegetables and fruits
- Red/purple/blue berries
- Other

The authors are currently evaluating the nutritional profiles of these groupings to determine their usefulness for researchers, dietitians, and nutrition educators to evaluate daily food choices, provide dietary guidance, and help consumers make daily selections of fruits and vegetables to meet nutrient and other food component needs.

4. Discussion

There are a variety of methods that can be used to conduct cluster analysis. This study reflects the method thought most appropriate for the nature of the database; however, several additional methods of cluster analysis were also explored. While it was found that cluster membership did shift with each variation in methodology, it was apparent that certain fruits and vegetables repeatedly grouped together regardless of the algorithm employed. Despite the variation of clustering results observed from different algorithms, this work suggests that there is value in employing mathematical clustering as a guide toward useful groupings.

4.1. Limitations of this study

This study has several limitations. The first is that some of the lesser-known food components such as glucosinolates and indole-3-carbinols were not included in the database and therefore could not influence cluster formation. Because these compounds are known to be found only in certain fruits and vegetables, some account was taken for this when proposing the groupings. Another limitation is that some data were imputed due to incomplete profiles. The degree to which this may have adversely affected cluster formation is not clear. Although this study encountered many of the difficulties inherent in a project where multiple measures are considered simultaneously, it was useful to examine fruit and vegetable classification in a new way. The results suggest that fruit and vegetable groupings based on combinations of botanic family, part of plant, color, and TAC may be predictive for some food components, but that other fruits and vegetables are unique in their composition and do not fit into neat groups.

4.2. Differences from MyPyramid

The fruit and vegetable classifications suggested in this work differ in several respects from those in the USDA MyPyramid (CNPP, 2008). First, the classifications suggested here are based on the composition of foods without regard for national consumption information. The MyPyramid vegetable groupings include considerations for typical (although not ideal) consumption patterns. Second, the classifications suggested here have a broader base in terms of the fruits and vegetables included and the food components under consideration. The MyPyramid food composites included 12 types of fruits and 36 different vegetables; the present study included 37 fruits and 67 vegetables. The present work also included evaluation of the presence of carotene, flavonoids, phytosterol, and TAC which were not considered in the MyPyramid work. Third, MyPyramid vegetable subgroups are intended to aid consumers, whereas the fruit and vegetable subgroups suggested here are geared to help researchers and other nutrition professionals as they evaluate food consumption data from surveys/studies and guide students, patients, and clients in appropriate food selections.

The MyPyramid website provides examples of fruits and vegetables that fit into the fruit group and the five vegetable subgroups; however, not all of these example vegetables were in the original composites (Tables 1 and 2). The placement of some fruits and vegetables within groupings is open to interpretation. MyPyramid places cherries within the "berry" group; this work does not consider cherries to be a type of berry. The MyPyramid clusters place olives within the cluster of peppers; this work does not consider olives to be a vegetable because they are highly processed and used more like condiments. MyPyramid places snow peas, asparagus, okra, and artichokes within the "green bean" cluster; this work does not consider these foods to have similar food composition levels. MyPyramid places tofu, soy-based meal replacements, and soy beverages with beans and peas; this work does not consider these foods to be vegetables. MyPyramid places green peas and green lima beans in the starchy vegetable group; this work considers these foods to be legumes. The green beans and green peppers, which are in the MyPyramid "other" group are about the same shade of green as broccoli, which is in the MyPyramid "dark green" group. The vegetable "mesclun" (listed in the MyPyramid "dark green" group) is not in SR20 (ARS, 2008). This term refers to a mix of young salad leaves such as lettuce, spinach, arugula (rocket), Swiss chard, mustard greens, endive, dandelion, etc.

5. Conclusions

The proposed classifications for fruits and vegetables are offered to nutrition professionals as a means to more accurately group fruits and vegetables based on food components of public health significance. The classifications may aid researchers who are developing food frequency questionnaires for epidemiological or clinical studies; they may be useful for instructors to teach students about food composition; and they may help dietitians provide dietary guidance to patients and clients.

Acknowledgements

The authors thank Susan Gebhardt and Joanne Holden of the USDA ARS Nutrient Data Laboratory for providing an electronic file of the fruit and vegetable data from the USDA Nutrient Database for Standard Reference, Release 20, which was used at the onset of this work.

References

Agricultural Research Service (ARS), United States Department of Agriculture (USDA), 2008. Nutrient Database for Standard Reference (SR), Release 20. Retrieved 2008-4-24 from: http://www.ars.usda.gov/Main/site_main.htm?modecode=12-35-45-00.

Akbay, A., Elhan, A., Özcan, C., Demirtas, S., 2000. Hierarchical cluster analysis as an approach for systematic grouping of diet constituents on basis of fatty acid, energy and cholesterol content: application on consumable lamb products. Medical Hypotheses 55, 147–154.

Britten, P., Marcoe, K., Yamini, S., Davis, C., 2006. Development of food intake patterns for the MyPyramid food guidance system. Journal of Nutrition Education and Behavior 38, S78–S92.

Cao, G., Sofie, E., Prior, R.L., 1996. Antioxidant capacity of tea and common vegetables. Journal of Agricultural and Food Chemistry 44, 3426–3431.

Center for Nutrition Policy and Promotion (CNPP), United States Department of Agriculture (USDA), 2008. MyPyramid Food Guide—Vegetable Subgroups. Retrieved 2008-4-24 from: http://www.mypyramid.gov/pyramid/vegetables_print.html.

Centers for Disease Control (CDC), Department of Health and Human Services, 2008. Fruits & Veggies—More Matters. Retrieved 2008-4-24 from: http://www.ars.usda.gov/Main/site_main.htm?modecode=12-35-45-00.

Dietary Guidelines Advisory Committee (DGAC), 2005. Report of the Dietary Guidelines Advisory Committee on the Dietary Guidelines for Americans, 2005. US Department of Agriculture, Agricultural Research Service. Document available from the National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22161; telephone 703 605-6000 and at http://www.health.gov/dietaryguidelines/dga2005/report/, pp. 3–4, 172–177.

Duester, K.D., 2001. Avocado fruit is a rich source of beta-sitosterol. Journal of the American Dietetic Association 101, 404–405.

Guo, C., Yand, J., Wei, J., Li, Y., Xu, J., Jiang, Y., 2003. Antioxidant activities of peel, pulp and seed fractions of common fruits as determined by FRAP assay. Nutrition Research 23, 1719–1726.

Lampe, J.W., 1999. Health effects of vegetables and fruit: Assessing mechanisms of action in human experimental studies. American Journal of Clinical Nutrition 70, 475S–490S.

Marcoe, K., Juan, W.Y., Yamini, S., Carlson, A., Britten, P., 2006. Development of food group composites and nutrient profiles for the MyPyramid food guidance system. Journal of Nutrition Education and Behavior 38, S93–S107.

Miller, H.E., Rigelhof, F., Marquart, L., Prakash, A., Kanter, M., 2000. Antioxidant content of whole grain breakfast cereals, fruits and vegetables. Journal of the American College of Nutrition 19, 312S–319S.

National Cancer Institute (NCI), National Institutes of Health, 1986. Diet, Nutrition & Cancer Prevention: The Good News. NIH Publication No. 87-2878, Bethesda, Maryland, US.

Norman, L., Johnsson, M., Andersson, H., van Gameren, Y., Dutta, P., 1999. Plant sterols in vegetables and fruits commonly consumed in Sweden. European Journal of Nutrition 38, 84–89.

Painter, J., Rah, J.H., Lee, Y.K., 2002. Comparison of international food guide pictorial representations. Journal of the American Dietetic Association 102, 483–489.

Pennington, J.A.T., 2003. Definitions and classifications for fruit and vegetables. In: IARC Handbooks of Cancer Prevention, volume 8, Fruit and Vegetables, IARC Press, Lyon, France, pp. 1–21.

Ruiz, D., Egea, J., Gil, M.I., Thomas-Barberan, F.A., 2005. Characterization and quantitation of phenolic compounds in new apricot (*Prunus armeniaca* L.) varieties. Journal of Agricultural and Food Chemistry 53, 9544–9552.

Ryan, E., Galvin, K., O'Connor, T.P., Maguire, A.R., O'Brien, N.M., 2007. Phytosterol, squalene, tocopherol content and fatty acid profile of selected seeds, grains, and legumes. Plant Foods and Human Nutrition 62, 85–91.

SAS Institute Inc., 2005. Applied Clustering Techniques. Instructor-based training. Cary, NC.

Schutz, K., Persike, M., Carle, R., Schieber, A., 2006. Characterization and quantification of anthocyanins in selected artichoke (*Cynara scolymus* L.) cultivars by HPLC-DAD-ESI-MS. Annals of Bioanalytical Chemistry 384, 1511–1517.

Souci, S.W., Fachmann, W., Kraut, H., 2000. Food Composition and Nutrition Tables, 6th ed. Medpharm Scientific Publishers/CRC Press, Stuttgart/ Washington DC.

Soong, Y.-Y., Barlow, P.J., 2004. Antioxidant activity and phenolic content of selected fruit seeds. Food Chemistry 88, 411–417.

Wang, H., Cao, G., Prior, R.L., 1996. Total antioxidant capacity of fruits. Journal of Agricultural and Food Chemistry 44, 701–705.

Weihrauch, J.L., Chatra, A.S., 1994. Provisional table on the vitamin K content of foods. USDA, ARS, HNIS (HNIS/PT-104), Washington DC.

Windham, C.T., Windham, M.P., Wyse, B.W., Hansen, R.G., 1985. Cluster analysis to improve food classification within commodity groups. American Dietetic Association 85, 1306–1314.

Wu, X., Beecher, G.R., Holden, J.M., Haytowitz, D.B., Gebhardt, S.E., Prior, R.L., 2004. Lipophilic and hydrophilic antioxidant capacities of common foods in the United States. Journal of Agricultural and Food Chemistry 52, 4026–4037.