نمایش داده‌های دو متغیره

نرگس سهرابی[۱]، هادی موقری[۲] ، کاظم فیاض‌حیدری[۳]

چکیده:

روش‌های عددی، اغلب برای یافتن پاسخِ سوال‌های از پیش تعیین شده مناسب هستند. در حالی که روش‌های گرافیکی محدودیت کمتری دارند. نمودارهای آماری، علاوه بر اینکه در شناسایی نقاط متمایز کاربرد دارند در درک روابط بین متغیرها نیز مفیدند. دستیابی به چنین اطلاعاتی توسط روش‌های عددی بسیار دشوار است. در این مقاله با استفاده از یک مجموعه داده‌ی واقعی، برخی از انواع نمایش داده‌های دو متغیره معرفی شده سپس یافته‌های حاصل با نتایج عددی مقایسه می‌شوند. همچنین بسته‌های نرم‌افزاری موجود در R برای رسم نمودارهای مورد مطالعه، معرفی خواهند شد.

واژه‌های کلیدی: داده‌های جفتی، داده‌های دو متغیره، نمودارهای آماری، نمودار کای، نمودار جعبه‌ای دو متغیره.

۱ مقدمه

هر تحلیل آماری از دو بخش تحلیل اکتشافی داده‌ها (EDA[۴]) و تحلیل تأئیدی داده‌ها (CDA[۵]) تشکیل می‌شود. EDA یک عملیات کارگاهی است به این معنا که آن با استفاده از روش‌های موجود به ویژه نمودارهای آماری، الگوهای بین مشاهدات را آشکار می‌کند. در مقابل CDA یک کار قضایی است که در آن شواهد موجود در داده‌ها به نفع یا علیه یک فرضیه، بررسی می‌شوند (دیگل و همکاران، ۱۹۹۴، صفحه‌ی ۳۳). به گفته‌ی توکی (۱۹۷۷)، EDA تمام ماجرا نیست با این حال چیز دیگری را هم نمی‌توان به عنوان اولین گام در تحلیل داده‌ها در نظر گرفت. تصویرسازی داده‌ها[۶] عبارت است از نمایش اطلاعات کمی و کیفی به گونه‌ای که بیننده بتواند الگوها، روندها و تغییرات را شناسایی کند. معمولاً مفهومی را که با دیدن یک تصویر درک می‌کنیم، بیش از مطالبی است که از راه خواندن به دست می‌آوریم. به همین دلیل است که گفته می شود: «یک عکس گویاتر از هزار کلمه است». در بسیاری از موارد، یک مجموعه داده‌ی آماری (حتی مجموعه‌های بزرگ) را می‌توان به خوبی توسط یک نمودار آماری تحلیل نمود و در موارد دیگر نیز نمودارهای آماری، نتایج بدست آمده از روش‌های عددی را به خوبی توصیف می‌کنند.

به طور کلی برای استفاده از نمودارهای آماری، می‌توان سه هدف عمده را برشمرد:

۱ - ثبت و ذخیره‌ی فشرده‌تر داده‌ها

۲ - ارایه‌ی اطلاعات به افراد دیگر

[۱] گروه آمار دانشگاه صنعتی امیرکبیر narges.sohrabi89@gmail.com

[۲] گروه آمار دانشگاه تربیت مدرس hadigilan@gmail.com

[۳] گروه آمار دانشگاه پیام‌نور – fayyaz@pnu.ac.ir

[4] Exploratory Data Analysis

[5] Confirmatory Data Analysis

[6] Data Visualization

۳- تحلیل داده‌ها

روش‌های عددی، اغلب برای یافتن پاسخِ سوال‌های از پیش تعیین شده مناسب هستند. در حالی که روش‌های گرافیکی محدودیت کمتری دارند. نمودارهای آماری، علاوه بر اینکه در شناسایی نقاط متمایز کاربرد دارند در درک روابط بین متغیرها نیز مفیدند. دستیابی به چنین اطلاعاتی توسط روش‌های عددی بسیار دشوار است (فدر، ۱۹۷۴).

نمودار دایره‌ای منسوب به ویلیام پلی‌فیر[7] معمولاً به عنوان اولین نمودار آماری شناخته می‌شود اما اخیراً تلاش صورت گرفته توسط ونلانگرن برای نمایش برآوردهایی از تفاوت طول جغرافیایی دو شهر تولدو[8] و رم به عنوان اولین نمودار آماری معرفی شده است (فرندلی و همکاران، ۲۰۱۰). تاریخچه‌ی استفاده از نمودارهای آماری و نمایش داده‌ها توسط آن‌ها بسیار طولانی است. برای بحث در این زمینه می‌توانید به وینر (۲۰۰۴) و فرندلی (۲۰۰۸b) مراجعه نمایید.

نمودارهای مدرن‌تر آماری در اوایل قرن ۱۸ام ظهور یافته و در بازه زمانی ۱۸۵۰ تا ۱۹۰۰ به طور گسترده‌ای در علوم مختلف به کار گرفته شدند. این دوران را دوران طلایی می‌نامند (فرندلی، ۲۰۰۸a). تا اوایل دهه‌ی ۱۹۷۰، استفاده از نمودارها در تحلیل داده‌ها در مقیاس وسیع امکان‌پذیر نبود زیرا اولاً امکانات نرم‌افزاری و سخت‌افزاری برای همه‌ی افراد مهیا نبوده ثانیاً رسم نمودارهای فراوان با استفاده از دست، زمانبر بودند. اما امروزه این مشکلات مرتفع شده و گسترش فن‌آوری، ایجاد نمودارهایی با کیفیت عالی را فراهم کرده است.

در این مقاله با استفاده از یک مجموعه داده‌ی واقعی، برخی از انواع مختلف نمایش داده‌های دو متغیره معرفی می‌شوند. این نمودارها شامل نمودار پراکندگی و روش‌های بهبود آن، نمایش داده‌های جفتی، نمودار کای و نمودار جعبه‌ای دو متغیره هستند. کلیه‌ی نمودارها با استفاده از نرم‌افزار R رسم می‌شوند. برای آشنایی با قابلیت‌های این نرم‌افزار در رسم نمودارهای آماری می‌توانید به مورل (۲۰۰۵) و میتال (۲۰۱۱) مراجعه نمایید.

۲ معرفی داده‌ها

داده‌های مورد استفاده در این تحقیق، مربوط به هزینه‌های بهداشت است و برخی از متغیرهای مرتبط با آن در ۴۱ کشور منتخب در حال توسعه از جمله ایران، جمع‌آوری شده است. داده‌های مذکور متعلق به بازه‌ی زمانی ۱۹۹۵ تا ۲۰۰۹ هستند.

متغیرهای موجود در این مجموعه داده عبارت از:

- hlte: هزینه‌ی بهداشت و سلامت
- gdp: تولید ناخالص داخلی
- lfex: میزان امید به زندگی
- ppaa: درصد جمعیت بالای ۶۵ سال
- mrtr: درصد مرگ و میر نوزادان
- urbp: درصد جمعیت شهرنشین

هستند.

۳ نمودار پراکندگی

پایه و اساس بسیاری از نمودارهای آماری در نمایش داده‌های دو متغیره، نمودار پراکندگی[9] است. این نمودار یکی از مفیدترین ابزارهای آماری برای تحلیل رابطه‌ی بین دو متغیر x و y است. فرض کنید $(x_i, y_i)\ i = 1, ..., n$ مشاهداتی از x و y باشند. نمودار y_i در برابر x_i به عنوان نمودار پراکندگی شناخته می‌شود.

در توصیف داده‌های دو متغیره، استفاده از یک شاخص آماری کفایت نمی‌کند. به عنوان مثال ضریب همبستگی گشتاوری پیرسون ممکن است همبستگی خطی بالایی را از نظر عددی بین x و y نشان دهد در حالی که بر اساس نمودار پراکندگی واقعاً بین این دو متغیر رابطه‌ی خطی وجود نداشته باشد و تنها به دلیل وجود مشاهدات پرت، ضریب همبستگیِ بزرگ بدست آمده باشد.

[7] William Playfair(1759-1823)

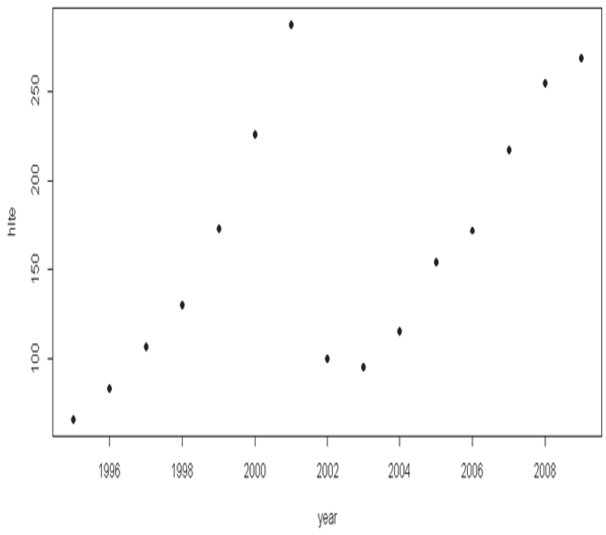
[8] Toledo: اسپانیا مرکز در واقع شهری

[9] Scatter Plot

[10] Factor- Response

چمبرز و همکاران (۱۹۸۳) نمودارهای پراکندگی را به دو گروه عمده تقسیم کردند. در گروه اول یکی از متغیرها (عامل) علت متغیر دیگر (پاسخ) است. این نوع را نمودار عامل- پاسخ[10] نامیدند. در رسم این نمودار، متغیر عامل روی محور افقی و متغیر پاسخ روی محور عمودی نشان داده می‌شود. هدف از رسم این نمودار، بررسی نحوه‌ی وابستگی متغیر پاسخ به متغیر عامل است. در گروه دوم هیچ یک از دو متغیر، علت متغیر دیگر نیست. این نوع را تعویض‌پذیر[11] نامیدند. در رسم این نمودار هر یک از دو متغیر را می‌توان روی محور افقی یا عمودی نشان داد. هدف از رسم این نمودار بررسی رابطه‌ی بین دو متغیر است.

نمودار سری زمانی یکی از نمونه‌های بارز نمودار پراکندگی از نوع عامل- پاسخ است. در این نمودار هدف، بررسی تاثیر زمان بر روی متغیر پاسخ است. در شکل ۱ سری زمانی هزینه‌های بهداشت و سلامت کشور ایران در بازه زمانی ۱۹۹۵ تا ۲۰۰۹ نشان داده شده است. همان گونه که از شکل پیداست، هزینه‌های بهداشت ایران از ابتدای بازه تا سال ۲۰۰۱ همواره صعود داشته اما در سال ۲۰۰۲ به یک باره کاهش یافته است. پس از آن دوباره سیر صعودی به خود گرفته است.



شکل ۱: نمودار سری زمانی هزینه‌ی بهداشت در ایران

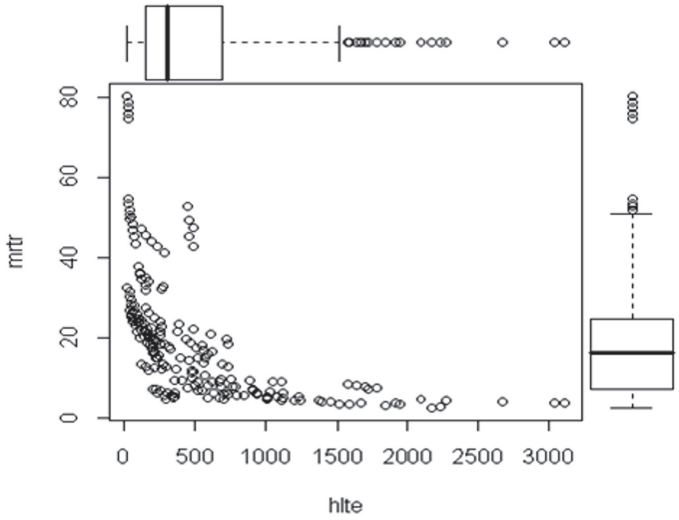
شایان ذکر است که برای آگاهی از پیشینه‌ی نمودار پراکندگی و چگونگی گسترش و تکمیل آن می‌توانید به فرندلی و دنیس (۲۰۰۵) مراجعه کنید.

۴ روش‌های بهبود نمودار پراکندگی

در این بخش برخی از روش‌های بهبود نمودار پراکندگی شامل افزودن نمودار جعبه‌ای و رفع همپوشانی نقاط را معرفی می‌کنیم. برای بحث تکمیلی در این زمینه می‌توانید به چمبرز و همکاران (۱۹۸۳، فصل چهار)، کلولند (۱۹۸۵، صفحه‌ی ۱۵۴) و ژاکویی (۱۹۹۷، فصل سه) مراجعه کنید.

۱.۴ افزودن نمودار جعبه‌ای

یکی از مرسوم‌ترین روش‌های بهبود نمودار پراکندگی، افزودن نمودار جعبه‌ای به حاشیه‌های این نمودار است. به این صورت که نمودار جعبه‌ای محور xها در بالای نمودار و نمودار جعبه‌ای محور yها در سمت راست نمودار رسم می‌شود. در شکل ۲ که نمودار پراکندگی درصد مرگ و میر نوزادان در برابر هزینه‌ی بهداشت و سلامت نشان داده شده است، نمودار جعبه‌ای این دو متغیر به حاشیه‌های نمودار پراکندگی افزوده شده است.



شکل ۲: نمودار پراکندگی درصد مرگ و میر نوزادان در برابر هزینه‌ی بهداشت به همراه نمودارهای جعبه‌ای

بر اساس این شکل با افزایش هزینه‌های بهداشت، نرخ مرگ و میر نوزادان به طور نمایی کاهش می‌یابد. علاوه بر این، نمودارهای جعبه‌ای هر دو متغیر نشان از وجود چندین مشاهده‌ی پرت در سمت راست توزیع می‌دهند.

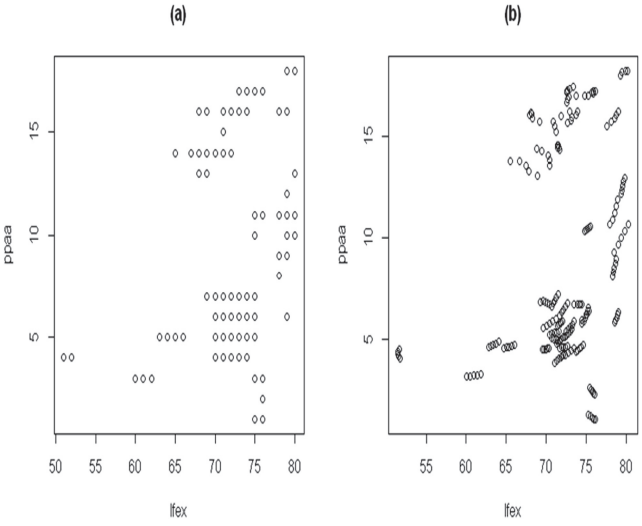
[11] Exchangeable

۲.۴ رفع همپوشانی نقاط

در بسیاری موارد پیش می‌آید که نقاط رسم شده در نمودار پراکندگی بر یکدیگر منطبق می شوند. این مساله مخصوصاً برای داده‌های گرد شده، رخ می‌دهد. در این صورت هر نقطه از نمودار پراکندگی ممکن است نماینده‌ی دو یا بیشتر از دو مشاهده باشد. در چنین حالتی، تفسیر نمودار پراکندگی به دشواری صورت می‌گیرد.

یکی از روش‌های رفع همپوشانی نقاط، افزودن عددی تصادفی به طول یا عرض نقاط و یا هر دوی آن‌ها است. به این ترتیب نقاط، اندکی جابه‌جا شده و در دستگاه مختصات متمایز می‌شوند. فرض کنید (x_i, y_i) مختصات نقطه $i-$ ام باشد. در این صورت مختصات نقطه‌ی جابه‌جا شده به صورت $(x_i + \theta_x u_i, y_i + \theta_y v_i)$ خواهد بود که در آن u_i و v_i(-۱، ۱). θ_x, θ_y نیز به ترتیب کسری از دامنه‌ی xها و yها هستند که معمولاً برابر ۰/۲۰ یا ۰/۵۰ در نظر گرفته می‌شوند. در نرم‌افزار R، برای جابه‌جایی نقاط می‌توان از تابع jitter استفاده نمود.

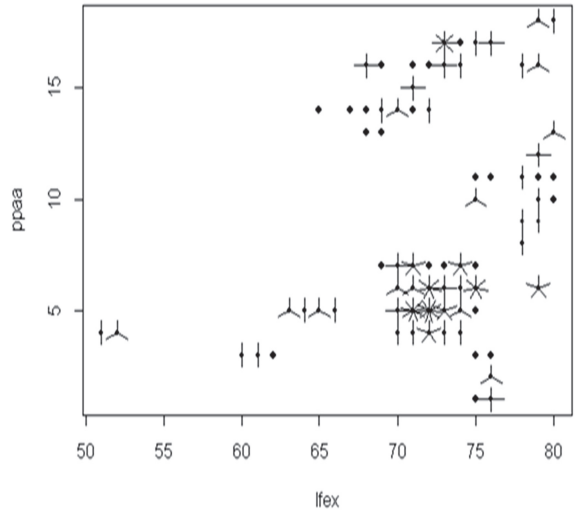
در شکل ۳ نمودار پراکندگی درصد جمعیت بالای ۶۵ سال در برابر تعداد سال‌های امید به زندگی در بازه‌ی زمانی ۲۰۰۵-۲۰۰۹ رسم شده است. در شکل ۳ (a)، داده‌های گردشده بسیار منظم به نظر می‌رسند و تنها ۷۶ نقطه قابل شناسایی است در صورتی که در مجموع، ۲۴۶ مشاهده رسم شده است. لذا برای آن‌که تأثیر جابه‌جایی نقاط بهتر نمایان شود، شکل ۳ (b) به صورتی که توضیح داده شد، رسم گردید.



شکل ۳: نمودار پراکندگی درصد جمعیت بالای ۶۵ سال در برابر امید به زندگی: (a) داده‌های گرد شده (b) داده‌های جابه‌جا شده

روش دیگر رفع همپوشانی نقاط در نمودار پراکندگی، استفاده از تکنیکی به نام گل آفتابگردان[12] است (کلولند و مک‌گیل، ۱۹۸۲). در این تکنیک تعداد مشاهدات به مرکزیت هر نقطه در نمودار با یک خط مشخص می‌شود. به این ترتیب یک نقطه‌ی تنها، نماینده‌ی یک مشاهده، یک نقطه به همراه دو خط نشان دهنده‌ی دو مشاهده است و الخ. برای استفاده از این تکنیک در محیط R تابع sunflowerplot تعبیه شده است. در شکل ۴ از تکنیک گل آفتابگردان برای رفع همپوشانی ایجاد شده در شکل ۳ (a) استفاده گردید.

شیلینگ و واتکینز (۱۹۹۴) با برشمردن چهار نقطه‌ی ضعف برای تکنیک گل آفتابگردان، فرم اصلاح شده‌ای از آن را ارائه نمودند.



شکل ۴: نمودار پراکندگی جمعیت بالای ۶۵ سال و امید به زندگی با استفاده از تکنیک گل‌آفتابگردان

۵ نمایش داده‌های جفتی

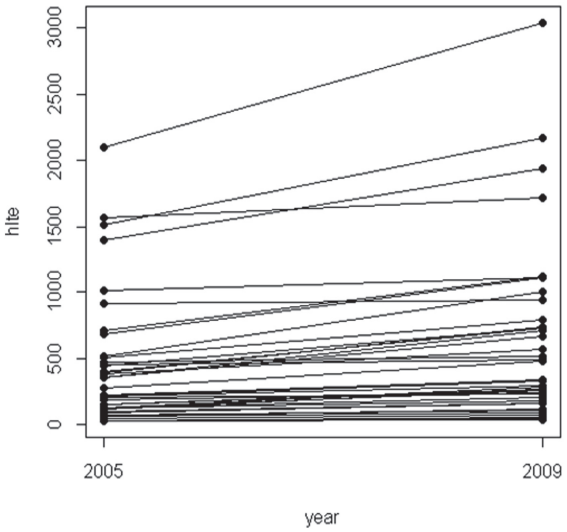
یک حالت خاص از داده‌های دو متغیره، داده‌های جفت‌شده[13] هستند. در این گونه داده‌ها که به اندازه‌های مکرر نیز شهرت دارند، هر آزمودنی دارای دو نوع اندازه‌ی پیش‌آزمون و پس‌آزمون است. چون آزمودنی‌ها به طور تصادفی و مستقل از یکدیگر انتخاب می‌شوند، هر زوج مشاهده، مستقل از زوج‌های دیگر است اما دو مشاهده‌ای که تشکیل یک زوج می‌دهند، مستقل از یکدیگر نیستند. برای آشنایی با مفاهیم مربوط به داده‌های جفتی و روش‌های تجزیه و تحلیل آن‌ها می‌توانید به بونت

[12] Sunflowers

[13] Paired Data

(۲۰۰۰) مراجعه کنید.

روشی که مخصوصاً در مجلات پزشکی برای نمایش داده‌های جفتی به کار برده می‌شود، عبارت از متصل نمودن داده‌های هر جفت نمودار موازی مختصات[14] معروف است. این روش به ویژه زمانی که حجم نمونه اندکی افزایش می‌یابد به دلیل همپوشانی داده‌ها، چندان مناسب نخواهد بود. به عنوان مثال فرض کنید، داده‌های هزینه‌ی بهداشت تنها حاوی اطلاعات دو سال ۲۰۰۵ و ۲۰۰۹ باشد در این صورت نمودار موازی مختصات همانند شکل ۵ خواهد شد.



شکل ۵: نمودار موازی مختصات هزینه‌ی بهداشت برای دو سال ۲۰۰۵ و ۲۰۰۹

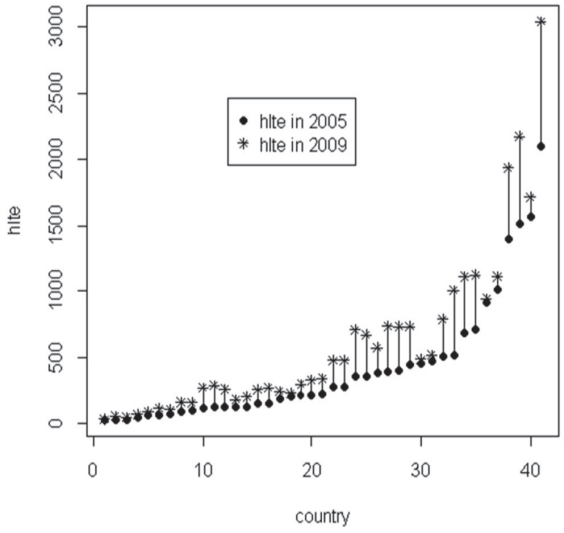
همان گونه که در شکل ۵ ملاحظه می‌شود، شناسایی الگوی تغییرات هزینه‌ی بهداشت از سال ۲۰۰۵ به سال ۲۰۰۹ برای ۴۱ کشور مورد مطالعه چندان قابل تشخیص نیست. مخصوصاً در بازه (۰,۵۰۰) که هزینه‌ی بهداشت اغلب کشورها در آن بازه قرار دارد.

یکی از راه‌کارهای مواجه با مساله‌ی همپوشانی در نمایش داده‌های جفتی استفاده از نمودار خطوط موازی[15] است (مک‌نیل، ۱۹۹۲). در این نمودار، ابتدا مشاهدات جفتی را بر اساس داده‌های پیش‌آزمون به طور صعودی مرتب می‌کنند سپس نمودار پراکندگی داده‌های پیش‌آزمون در برابر ترتیب مشاهدات رسم می‌شود. در نهایت داده‌های پس‌آزمون توسط خطوط عمودی به داده‌های پیش‌آزمون متناظر متصل می‌شوند.

شکل ۶، نمودار خطوط موازی را برای داده‌های هزینه‌ی بهداشت در دو سال ۲۰۰۵ و ۲۰۰۹ نشان می‌دهد. در این شکل برخلاف نمودار موازی مختصات (شکل ۵)، الگوی کلی تغییراتِ هزینه‌ی بهداشت در بین کشورهای مورد مطالعه به خوبی قابل تشخیص است. از نمودار مذکور دو نکته برداشت می‌شود:

۱. هزینه‌ی بهداشت تمام کشورها از سال ۲۰۰۵ به سال ۲۰۰۹، افزایش داشته است.

۲. به طور کلی میزان افزایش هزینه های بهداشت کشورهایی که دارای هزینه‌ی بهداشت بیشتری در سال ۲۰۰۵ بوده‌اند، بیش از میزان افزایش هزینه‌های بهداشت سایر کشورها است.



شکل ۶: نمودار خطوط موازی هزینه‌ی بهداشت برای دو سال ۲۰۰۵ و ۲۰۰۹

۶ نمودار جعبه‌ای دو متغیره

نمودار جعبه‌ای یا نمودار جعبه و خط[16] یکی از بهترین ابزارها برای نمایش خلاصه‌ی داده‌ها است. در این نمودار که اولین بار توسط توکی (۱۹۷۷) معرفی گردید، پنج آماره‌ی مینیمم، چارک اول، میانه، چارک سوم و ماکسیمم نمایش داده می‌شود.

از زمان معرفی نمودار جعبه‌ای تا کنون انواع مختلفی از آن ارائه شده است (به عنوان مثال به هینتز و نلسن، ۱۹۹۸، کمپ‌استرا، ۲۰۰۸ و هوبرت و وندرویرن، ۲۰۰۸، مراجعه شود). نمودار جعبه‌ای دو متغیره[17] یکی

[14] Parallel Coordinate Plot

[15] Paralle Line Plot

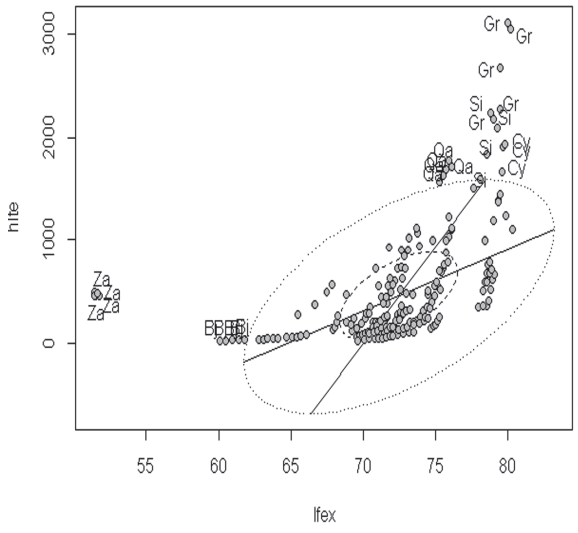
[16] Box-and-Whisker Plot

[17] Bivariate Boxplot

از مفیدترین تعمیم‌هایی است که بر نمودار جعبه‌ای تک متغیره صورت گرفته است (گلدبرگ و ایگلویچ، ۱۹۹۲). این نمودار، مخصوصاً برای شناسایی نقاط پرت و قضاوت در مورد نرمال بودن توزیع توأم مناسب است.

نمودار جعبه‌ای دو متغیره اساساً از دو بیضی هم مرکز تشکیل می‌شود. بیضی درونی تحت عنوان لولا، ۵۰ درصد کل نقاط را در بر می‌گیرد و بیضی بیرونی تحت عنوان حصار[18] نقاط پرت را مشخص می‌کند. علاوه بر این، خط رگرسیون x روی y و رگرسیون y روی x در این نمودار رسم می‌شود. در صورتی که دو متغیر، شدیداً همبسته باشند (مستقیم یا معکوس) زاویه‌ی تند بین خطوط رگرسیونی کوچک خواهد شد و در مقابل برای همبستگی‌های کوچک، این زاویه بزرگ می‌شود.

تابع bv.boxplot از بسته نرم‌افزاری asbio برای رسم نمودار جعبه‌ای دو متغیره در محیط R طراحی شده است. شکل ۷ نمودار جعبه‌ای دو متغیره‌ی هزینه‌ی بهداشت (hlte) و امید به زندگی (lfex) را نشان می‌دهد. همبستگی پیرسون بین این دو متغیر برابر ۰/۵۲ است. مشاهدات مربوط به کشورهای آفریقای جنوبی، بنین، یونان، قطر، قبرس و اسلونی که در شکل به ترتیب با حروف Za، Bj، Gr، Qa، Cy، Si نشان داده شده‌اند، دور افتاده محسوب می‌شوند، زیرا مشاهدات این کشورها خارج از حصار قرار گرفته‌اند.



شکل ۷: نمودار جعبه‌ای دو متغیره هزینه‌ی بهداشت و امید به زندگی

۷ نمودار کای

علی‌رغم کاربرد گسترده‌ی نمودار پراکندگی در ارزیابی وابستگی بین دو متغیر پیوسته، اغلب قضاوت در مورد وجود وابستگی به دلیل الگوی خاص پراکندگی نقاط، مشکل است. فیشر و سوئیتزر (۱۹۸۵ و ۲۰۰۱) برای تسهیل این امر، نموداری را تحت عنوان نمودار کای[19] معرفی کرده‌اند. این نمودار در واقع نمودار پراکندگی زوج‌های (λ_i, χ_i) است که در آن:

$$\chi_i = \frac{H_i - F_i G_i}{\{F_i(1-F_i)G_i(1-G_i)\}^{1/2}}$$

$$\lambda_i = 4S_i \max\left\{(F_i - \tfrac{1}{2})^2, (G_i - \tfrac{1}{2})^2\right\},$$

و

$$H_i = \frac{1}{n-1}\sum_{i\neq j} I(x_j \le x_i, y_j \le y_i)$$

$$F_i = \frac{1}{n-1}\sum_{i\neq j} I(x_j \le x_i)$$

$$G_i = \frac{1}{n-1}\sum_{i\neq j} I(y_j \le y_i)$$

$$S_i = sign\Big\{(F_i - \tfrac{1}{2}).(G_i - \tfrac{1}{2})\Big\}.$$

در روابط اخیر، $sign(.)$ و $I(.)$ به ترتیب نشان دهنده‌ی تابع علامت و تابع نشانگر هستند.

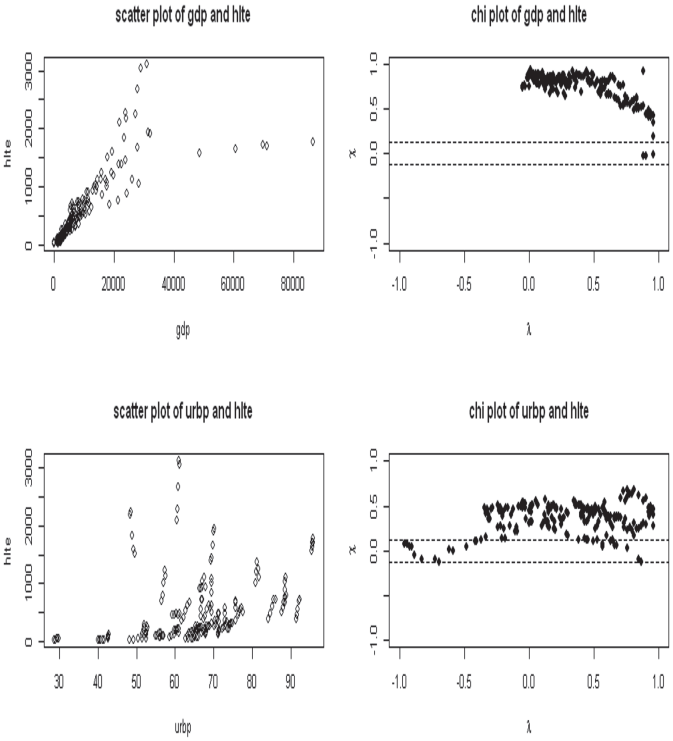
هنگامی که دو متغیر مورد مطالعه مستقل باشند، نقاط رسم شده در نمودار کای، تقریباً حول محور xها قرار می‌گیرند. در صورتی که نقاط خارج از این محدوده قرار گیرند، به معنای وجود وابستگی بین دو متغیر است. برای رسم نمودار کای در محیط R تابع chiplot از بسته‌ی نرم‌افزاری MVA قابل استفاده است.

نمودار کای برای زوج‌های (gdp, hlte) و (urbp, hlte) در شکل ۸ نشان داده شده است. علاوه بر این، نمودار پراکندگی این دو جفت متغیر نیز رسم شده است. بر اساس نمودارهای پراکندگی به نظر می‌رسد بین دو متغیر hlte و gdp یک رابطه‌ی قوی خطی برقرار است. اما قضاوت در مورد نمودار پراکندگی دو متغیر urbp و hlte اندکی مشکل است. نقاط رسم شده در نمودار کای مربوط به دو متغیر gdp و hlte نسبت به نقاط نمودار کای مربوط به دو متغیر urbp و hlte تمایل بیشتری دارند که از محور xها فاصله بگیرند و تنها تعداد اندکی از آن‌ها بین دو خط موازی مشخص شده در شکل قرار گرفته‌اند. در حالی که در نمودار کای متعلق

[18] Fence

[19] Chiplot

به دو متغیر urbp و hlte نقاط به محور xها نزدیک‌تر هستند و تعداد بیشتری از آن‌ها بین دو خط موازی قرار دارند.



شکل ۸: نمودار پراکندگی هزینه‌ی بهداشت در برابر تولید ناخالص داخلی و درصد جمعیت شهرنشین (سمت چپ) و نمودارهای کای متناظر (سمت راست)

بیان این نکته ضروری است که نمودار دیگری تحت عنوان نمودار کندال[20] توسط گنست و بویس (۲۰۰۳) برای تشخیص همبستگی بین دو متغیر ابداع شده است.

۸ یافته‌های عددی

در این بخش با استفاده از مدل مولفه‌ی خطای تک عاملی[21] با اثرات تصادفی، به بررسی ارتباط بین دو متغیر هزینه‌ی بهداشت (hlte) و امید به زندگی (lfex) می‌پردازیم.

دلیل استفاده از مدل مولفه‌ی خطا، ساختار ویژه‌ی داده‌های مورد بررسی است. مجموعه‌ی داده‌های هزینه‌ی بهداشت، شامل اطلاعات ۴۱ کشور (بعد مقطع) در بازه زمانی ۱۹۹۵ تا ۲۰۰۹ (بعد زمان) است. چنین داده‌هایی را داده‌های طولی[22] می‌نامند که در اقتصاد و جامعه‌شناسی با عنوان داده‌های پانلی[23] شناخته می‌شوند (بالتاگی، ۲۰۰۵ و دیگل و همکاران، ۱۹۹۴). وجود همبستگی بین داده‌ها، ایجاب می‌کند که از مدلی برای تحلیل داده‌های طولی استفاده شود که امکان مدل‌سازی این همبستگی را فراهم نماید. مدل مذکور، با وارد کردن اثر تصادفی واحدهای مقطعی این امکان را فراهم می‌آورد.

مدل مولفه‌های خطا با اثرات تصادفی به صورت زیر است:

$$\ln hlte_{it} = \beta_0 + \beta_1 lfex_{it} + b_i + e_{it}$$

که در آن $b_i \sim N(0, \sigma_b^2)$ اثر تصادفی واحدهای مقطعی (کشورها) و $e_{it} \sim N(0, \sigma_e^2)$ جمله‌ی خطا است. در مدل فوق فرض می‌شود که دو مولفه‌ی b_i, e_{it} از یکدیگر مستقل هستند. لازم به ذکر است که به دلیل چوله بودن توزیع هزینه‌ی بهداشت از لگاریتم آن به عنوان متغیر پاسخ استفاده شده است. نتایج حاصل از برآورد مدل رگرسیونی در جدول ۱ نشان داده شده است. برای برآورد مدل، تابع lme از بسته‌ی نرم‌افزاری nlme به کار گرفته شد. بر اساس نتایج بدست آمده، با افزایش امید به زندگی، هزینه‌های بهداشت و سلامت نیز افزایش می‌یابد.

اینک می‌خواهیم با استفاده از تحلیل تشخیصی[24]، صحت و سقم یافته‌های بدست آمده از نمودار جعبه‌ای دو متغیره (به عنوان نمونه) را بررسی کنیم. برای تعیین مشاهدات پرت و تاثیرگذار[25] در چارچوب داده‌های طولی، معمولاً از معیارهای زیر استفاده می‌شود (لیتل و همکاران، ۲۰۰۶، صفحه‌ی ۴۱۳):

$$\begin{aligned}
L_j &= 2\Big\{\ell(\hat{\psi}) - \ell(\hat{\psi}_{(j)})\Big\}, \\
D(\beta) &= (\hat{\beta} - \hat{\beta}_{(j)})'[\hat{Var}(\hat{\beta})]^{-1} \\
&\quad (\hat{\beta} - \hat{\beta}_{(j)})/rank(\mathbf{X}), \\
COVRATIO(\beta) &= \frac{det_{ns}\Big(\widehat{Var}(\hat{\beta}_{(j)})\Big)}{det_{ns}\Big(\widehat{Var}(\hat{\beta})\Big)}, \\
PRESS_{(j)} &= \sum_{k \in j}(y_{it} - \mathbf{x}'_{\mathbf{it}}\hat{\beta}_{(\mathbf{j})})^2.
\end{aligned}$$

[20] Kendall Plot

[21] One-Way Error Component Model

[22] Longitudinal Data

[23] Panel Data

[24] Diagnostic Analysis

[25] Influential Observation

در معیار اول، $\hat{\psi}$ بردار برآورد پارامترهای موجود در مدل و $\ell(\hat{\psi})$ لگاریتم تابع درستنمایی به ازای $\hat{\psi}$ است. به طور متناظر $\ell(\hat{\psi}_{(j)})$، لگاریتم تابع درستنمایی به ازای $\hat{\psi}$ پس از حذف مشاهدات واحد $j-$ام می‌باشد. در معیار دوم، $rank(\mathbf{X})$ رتبه‌ی ماتریس مشاهدات $\mathbf{X}$ است. همچنین $\hat{\beta}_{(j)}$ برآورد بردار β است که بدون در نظر گرفتن واحد $j-$ام بدست می‌آید. در معیار سوم، $det_{ns}(\mathbf{M})$ نشان دهنده‌ی دترمینان بخش ناتکین ماتریس $\mathbf{M}$ می‌باشد.

معیار L_j تاثیر کل واحد $j-$ام را اندازه‌گیری می‌کند. معیار $D(\beta)$ و $COVRATIO(\beta)$ به ترتیب برای تشخیص مشاهدات تاثیرگذار بر برآورد و دقت برآورد اثرات ثابت به کار برده می‌شوند. بالاخره معیار $PRESS_j$ برای شناسایی مشاهدات موثر بر مقادیر برازش شده کاربرد دارد. مقادیر بزرگ L_j، $D(\beta)$ و $PRESS_j$ و نیز مقادیر کوچک $COVRATIO(\beta)$ به منزله تاثیرگذار بودن واحد مربوطه بر پارامتر مورد نظر است.

بر طبق جستجوهای صورت گرفته در بین نرم‌افزارهای موجود، تنها در نرم‌افزار SAS امکاناتی برای محاسبه شاخص‌های فوق در چارچوب داده‌های طولی فراهم شده است (لیتل و همکاران، ۲۰۰۶). در جدول ۲، مقادیر شاخص‌های فوق به ازای هر یک از کشورهای مورد مطالعه ارائه شده است.

همان گونه که در جدول ۲ ملاحظه می‌شود، معیارهای متناظر با کشور آفریقای جنوبی با مقادیر متناظر سایر کشورها تفاوت بسیاری دارد و این دلالت بر موثر بودن مشاهدات کشور مذکور بر برآورد پارامترهای مدل دارد. این نتیجه با یافته‌ی بدست آمده از نمودار جعبه‌ای دو متغیره مطابقت می‌کند. بر اساس نمودار مذکور (شکل ۷)، مشاهدات کشور آفریقای جنوبی به عنوان مشاهدات پرت شناسایی شده بودند.

۹ بحث و نتیجه‌گیری

اولین گام در تحلیل داده‌ها، استفاده از نمودارهایی است که اطلاعات و الگوهای بین مشاهدات را شناسایی کرده و به بهترین وجه ممکن به نمایش گذارد. نکته‌ی حائز اهمیت در این بین، انتخاب نمودار آماری مناسب است. امروزه استفاده از نمودارهای آماری چنان گسترش یافته است که در مجامع علمی به طور ویژه‌ای به آن پرداخته می‌شود و نویسندگان مختلف قوانین و قواعد خاصی را برای نمودارهای آماری وضع کرده‌اند (کلولند، ۱۹۸۵ و ویلکینسن، ۲۰۰۵).

در این مقاله برخی از روش‌های نمایش داده‌های دو متغیره معرفی شدند. نمودار پراکندگی، پایه و اساس این نمودارها است. برخی از روش‌های بهبود این نمودار نظیر افزودن نمودار جعبه‌ای و رفع همپوشانی نقاط، بیان گردید. دو روش متداول برای نمایش داده‌های جفتی نیز معرفی شدند. علاوه بر این، به نمودارهای جدیدتر شامل نمودار کای و نمودار جعبه‌ای دو متغیره نیز پرداخته شد.

تشکر و قدردانی

نویسندگان مقاله از آقای مهدی باسخا، به دلیل فراهم نمودن داده‌های هزینه‌ی بهداشت و سلامت، کمال تشکر را دارند. همچنین از داوران محترم مقاله به خاطر نظرات ارزشمند در جهت بهبود ساختار مقاله، تشکر و قدردانی می‌شود.

جدول ۱: نتایج حاصل از برآورد مدل مؤلفه‌ی خطا با اثرات تصادفی

متغیر	ضریب	انحراف معیار
ضریب ثابت	-۱۲/۷۹	۱/۵۰
lfex	۰/۲۶	۰/۰۲
b	۱/۱۷	
e	۰/۱۶	

جدول ۲: تحلیل تشخیصی داده‌های هزینه‌ی بهداشت و سلامت

کشور	L_j	$PRESS_{(j)}$	$D()$	$COVRATIO()$
آرژانتین	۰۳۰/۰	۲۳۳/۰	۰۰۹/۰	۰۴۴/۱
آذربایجان	۱۳۵/۰	/۲۶۹	۰۵۵/۰	۰۱۴/۱
بحرین	۰۲۷/۰	۲۳۵/۰	۰۰۸/۰	۰۵۵/۱
بلورس	۰۴۹/۰	۲۹۳/۱	۰۰۲/۰	۱۰۶/۱
بنین	۱۹۵/۰	۴۲۹/۲	۱۲۷/۰	۲۲۱/۱
بولیوی	۰۵۸/۰	۳۹۰/۰	۰۰۱/۰	۱۱۲/۱
برزیل	۰۵۱/۰	۳۵۶/۲	۰۳۱/۰	۰۲۳/۱
بلغارستان	۰۲۷/۰	۱۵۲/۰	۰۰۵/۰	۰۴۸/۱
شیلی	۰۱۹/۰	۷۷۲/۳	۰۰۷/۰	۰۶۴/۱
کلمبیا	۰۴۱/۰	۳۷۷/۰	۰۰۹/۰	۰۶۰/۱
کرواسی	۰۱۹/۰	۴۲۷/۰	۰۰۳/۰	۰۵۷/۱
کاستاریکا	۰۱۹/۰	۵۷۳/۷	۰۱۴/۰	۰۳۱/۱
کوبا	۱۳۹/۰	۱۱۵/۶	۰۱۷/۰	۰۱۱/۱
قبرس	۰۲۷/۰	۲۱۴/۰	۰۰۰/۰	۰۸۷/۱
اکوادر	۰۳۳/۰	۲۶۸/۷	۰۱۸/۰	۰۱۶/۱
مصر	۰۲۷/۰	۷۶۲/۲	۰۳۰/۰	۰۲۵/۱
السالوادر	۰۶۰/۰	۰۲۵/۰	۰۰۱/۰	۰۹۳/۱
استونی	۰۳۹/۰	۴۶۹/۰	۰۱۱/۰	۱۲۸/۱
گرجستان	۳۷۵/۰	۸۱۳/۰	۰۱۰/۰	۰۳۳/۱
یونان	۰۴۳/۰	۴۱۴/۰	۰۰۸/۰	۰۸۱/۱
مجارستان	۰۵۲/۰	۲۱۵/۴	۰۲۷/۰	۰۹۶/۱
هند	۰۴۳/۰	۲۲۰/۰	۰۰۱/۰	۱۱۲/۱
اندونزی	۲۳۶/۰	۳۶۰/۱۲	۱۵۷/۰	۹۴۹/۰
ایران	۰۴۲/۰	۱۴۹/۰	۰۲۰/۰	۰۴۷/۱
کره جنوبی	۴۱۸/۰	۹۴۵/۱	۳۰۰/۰	۲۲۲/۱
لیتوانی	۵۴۴/۰	۹۹۹/۴	۲۲۶/۰	۱۵۰/۱
مالزی	۰۳۴/۰	۵۲۶/۱	۰۰۸/۰	۰۵۶/۱
مکزیک	۰۴۹/۰	۱۱۱/۰	۰۰۷/۰	۱۰۳/۱
مراکش	۰۶۲/۰	۷۳۵/۱	۰۱۹/۰	۰۵۵/۱
نیکاراگوئه	۰۹۴/۰	۶۹۳/۹	۰۲۹/۰	۰۳۰/۱
پاراگوئه	۰۶۰/۰	۳۰۳/۳	۰۲۵/۰	۰۱۳/۱
پرو	۰۵۲/۰	۵۳۹/۳	۰۲۲/۰	۰۳۷/۱
فیلیپین	۱۴۵/۰	۴۲۱/۱۲	۰۷۹/۰	۹۶۹/۰
قطر	۰۵۸/۰	۵۰۱/۳	۰۰۷/۰	۰۶۵/۱
روسیه	۵۰۶/۰	۱۲۶/۱۵	۳۶۰/۰	۲۵۹/۱
اسلونی	۰۵۳/۰	۳۱۰/۰	۰۰۵/۰	۰۹۵/۱
آفریقای جنوبی	۳۲۴/۲۸	۱۸۸/۱۹۹	۸۵۴/۰	۳۸۵/۰
تونس	۰۵۷/۰	۵۰۵/۳	۰۰۷/۰	۰۶۲/۱
ترکیه	۰۲۱/۰	۲۲۶/۲	۰۰۸/۰	۰۴۴/۱
اکراین	۴۴۶/۰	۷۹۲/۱	۱۲۰/۰	۱۰۵/۱
الجزایر	۳۴۸/۰	۵۵۸/۱	۰۳۱/۰	۹۹۰/۰

مراجع

[1] Baltagi, B. H. (2005). *Econometric Analysis of Panel Data*, 3rd ed, John Wiles and Sons: Chichester.

[2] Bonate, P. L. (2000). *Analysis of Pretest-Posttest Designs*, Chapman & Hall/CRC.

[3] Cleveland, W. S,(1985). *The Elements of Graphing Data*, Montery.

[4] Cleveland, W. S. and McGill, R. (1984). The Many Faces of a Scatterplot, *Journal of the American Statistical Association*, **79**, 807-822.

[5] Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994). *Analysis of Longitudinal Data*, Oxford University press.

[6] Feder, P. I. (1974). Graphical Techniques in Statistical Data Analysis-Tools for Extracting Information from Data, *Technometrics*, **16**, 287-299.

[7] Fisher, N. I. and Switzer, P. (1985). Chi-Plots for Assessing Dependence, *Biometrika*, **72**, 253-265.

[8] Fisher, N. I., and Switzer, P. (2001). Graphical Assessment of Dependence, *The American Statistician*, **55**, 233-239.

[9] Friendly, M. (2008a). The Golden Age of Statistical Graphics, *Statistical Science*, **23**, 502-535.

[10] Friendly, M. (2008b). A Brief History of Data Visualization, *Handbook of Data Visualization*, 15-56.

[11] Friendly, M. and Denis, D. (2005). The Early Origins and Development of the Scatterplot, *Journal of the History of the Behavioral Sciences*, **41**, 103-130.

[12] Friendly, M., Valero-Mora, P. and Ulargui, J. I. (2010). The First (Known) Statistical Graph: Michael Florent van Langren and the ”Secret” of Longitude, *The American Statistician*, **64**, 174-184.

[13] Genest, C., and Boies, J. C. (2003). Detecting Dependence with Kendall Plots, *The American Statistician*, **57**, 275-284.

[14] Goldberg, K. M. and Iglewicz, B. (1992). Bivariate Extensions of the Boxplot, *Technometrics*, **34**, 307-320.

[15] Hintze, J. L. and Nelson, R. D. (1998). Violin Plots: a Box Plot-Density Trace Synergism, *The American Statistician*, **52**, 181-184.

[16] Hubert, M. and Vandervieren, E. (2008). An Adjusted Boxplot for Skewed Distributions, *Computational Statistics & Data Analysis*, **52**, 5186-5201.

[17] Jacoby, W. G. (1997). *Statistical Graphics for Univariate and Bivariate Data*, Sage Publications, Incorporated.

[18] Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions, *Journal of Statistical Software*, **28**.

[19] Littel, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D. and Schabenberger, O. (2006). *SAS for Mixed Models*, 2nd ed. SAS Publishing, Cary, NC.

[20] McNeil, D. (1992). On Graphing Paired Data, *The American Statistician*, **46**, 307-311.

[21] Mittal, H. V. (2011). *R Graph Cookbook*, Packt Pub Limited.

[22] Murrell, P. (2005). *R Graphics*. Chapman & Hall/CRC.

[23] Schilling, M. F. and Watkins, A. E. (1994). A Suggestion for Sunflower Plots, *The American Statistician*, **48**, 303-305.

[24] Tukey, J. W. (1977). *Exploratory Data Analysis*, Reading: Addison-Wesley.

[25] Wainer, H. (2004). *Graphic Discovery: A Trout in the Milk and other Visual Adventures*. Princeton University Press.

[26] Wilkinson, L. (2005). *The Grammar of Graphics*. Springer.