کاربرد تحلیل خوشه‌ای و تحلیل عاملی در مطالعات اقتصادی و محیط‌زیستی چندمقطعی با استفاده از مؤلفه‌های مؤثر

وحید ماجد[1]، حسین میرشجاعیان حسینی[2]، سمیرا ریاضی دوست[3]

تاریخ دریافت: ۹۷/۳/۱٤ تاریخ پذیرش: ۹۷/۱۲/۱۸

چکیده

همگنی گروه‌ها در مطالعاتی که از روش‌های چند مقطعی استفاده می‌نمایند از اهمیت ویژه‌ای برخوردار است. این امر در مطالعات اقتصادی بویژه مطالعات متکی بر روش داده‌های تابلویی برای اعتبار نتایج و برآوردها اهمیت زیادی دارد. در تحلیل‌های چند مقطعی با مقاطع زیاد، بویژه تحلیل‌های مبتنی بر داده‌های تابلویی، خوشه‌بندی ضمن افزایش اطمینان از همگنی موردنظر و استحکام و اعتبار نتایج بدست‌آمده، امکان مقایسه گروه‌های مختلف با ویژگی‌های متفاوت را نیز فراهم می‌آورد. در این مقاله به ارائه روش‌های مرسوم در خوشه‌بندی و همگن‌سازی گروه‌ها در مطالعات چند مقطعی اقتصادی و محیط‌زیستی بر مبنای مؤلفه‌های مؤثر پرداخته شده است. بدین منظور نمونه‌ای متشکل از ۹۲ کشور با بیشترین میزان انتشار CO_2 در دوره زمانی ۱۹۹۰ تا ۲۰۱۲ که داده‌های مربوط به آنها در این دوره در دسترس بوده است، براساس ۱۸ معیار مؤثر خوشه‌بندی شده‌اند. معیارهای مذکور با استفاده از تحلیل عاملی به پنج مؤلفه اصلی تقلیل پیدا کرده و خوشه‌بندی کشورها به روش سلسله مراتبی بر مبنای مؤلفه‌های اصلی (HCPC) انجام شده است. انجام خوشه‌بندی به تفکیک ۹۲ کشور به هفت خوشه متفاوت هرکدام با ویژگی‌های خاص منجر شده است. بررسی مشخصات غالب کشورها، نشان از همگنی در هر یک از خوشه‌های مشخص‌شده دارد.

واژه‌های کلیدی: تحلیل خوشه‌ای، تحلیل عاملی، مطالعات چندمقطعی، مؤلفه‌های اقتصادی، مؤلفه‌های محیط‌زیستی.

طبقه‌بندی JEL: Q54, C38, C33.

---

۱. استادیار دانشکده اقتصاد دانشگاه تهران (نویسنده مسئول) Email:Majed@ut.ac.ir

۲. استادیار پژوهشکده امور اقتصادی وزارت امور اقتصادی و دارایی Email:hmirshojaeian@gmail.com

۳. کارشناس ارشد دانشکده اقتصاد دانشگاه تهران Email:s.riazi@ut.ac.ir

۱. مقدمه

در تحلیل‌های آماری و اقتصادی، خوشه‌بندی[1] صحیح گروه‌های مختلف کشورها یا بنگاه‌های مورد بررسی از اهمیت زیادی برخوردار است. به‌ویژه در تحلیل‌های اقتصادسنجی مرسوم که از مدل داده‌های تابلویی[2] استفاده می‌نمایند، همگنی کامل یا جزئی پارامترها برای تخمین مدل ضروری است. از سوی دیگر انتخاب گروهی از موارد ناهمگن یا گروه‌بندی نادرست آن‌ها در تحلیل داده‌های تابلویی، نه‌تنها برآوردهای نادرستی را نتیجه می‌دهد، بلکه باعث می‌شود آزمون‌ها نیز ناسازگار شوند. بنابراین انتخاب مناسب یک گروه یا گروه‌های مختلف برای ساختن یک مدل داده‌های تابلویی (با توجه به موضوع موردمطالعه) بسیار مهم است (استلیاروا[3]، ۲۰۱۳). براین اساس مطالعات اقتصادسنجی با داده‌های تابلویی که تاکنون انجام گرفته‌اند، تلاش کرده‌اند برمبنای یک یا چند معیار مختلف، کشورها را در گروه‌هایی همگن گردآوری کنند. اغلب این معیارها، (بدون توجه به حوزه خاص مورد مطالعه) سطح توسعه اقتصادی، طبقه درآمدی، وابستگی منطقه‌ای و گروه‌های متحد اقتصادی هستند. به‌طورخاص در مطالعات محیط‌زیستی می‌توان به نمونه‌های زیر اشاره نمود: هانگ، هوانگ و یانگ[4](۲۰۰۸) تعداد ۸۲ کشور را به چهار گروه درآمدی طبقه‌بندی کردند و نیکواقبال وهمکاران (۱۳۹۱) تحلیلی مبتنی بر داده های تابلویی بر روی سه گروه درآمدی انجام دادند. ولد–رافائل[5] (۲۰۰۶)، لی و چانگ[6] (۲۰۰۷)، اسو[7] (۲۰۱۰) و پورکاظمی و ابراهیمی (۱۳۸۷) کشورها را براساس منطقه گردآوری کردند، همچنین دیندا و کوندو[8] (۲۰۰۶)، کشورهای موردمطالعه خود را به سه گروه منطقه‌ای متفاوت تفکیک نمودند. برخی مطالعات نیز بر گروه‌های متحد اقتصادی یا با سطح یکسان توسعه اقتصادی تمرکز کردند. به عنوان مثال اسکوالی[9] (۲۰۰۶) کشورهای عضو OPEC[1]،

1. Clustering
2. Panel Data
3. Stolyarova
4. Huang, Hwang & Yang
5. Wolde-Rufael
6. Lee & Chang
7. ESSO
8. Dinda & Coondoo
9. Squali

بلک[۲] و همکاران (۲۰۱۰)، ۲۵ کشور عضو OECD[۳] و پائو و تسای[۴] (۲۰۱۰) کشورهای BRICs[۵] را مطالعه نمودند. اما آنچه مسلم است با توجه به موضوع موردمطالعه معیارهای طبقه‌بندی باید متفاوت باشد. بنابراین در مطالعات محیط‌زیستی، گروه‌بندی تنها براساس تقسیم منطقه‌ای یا سطح توسعه اقتصادی، کافی به نظر نمی‌رسد و می‌تواند به تخمینی نادرست از پارامترها منجر شود و قدرت آزمون‌ها را نیز تضعیف نماید.

به‌طورخاص درمورد مطالعاتی که به بررسی عوامل مؤثر بر انتشار گازهای گلخانه‌ای مانند رشد اقتصادی و مصرف انرژی می‌پردازند، می‌توان چنین استدلال کرد که کشورهایی با سطح یکسان توسعه اقتصادی به دلایل بسیاری ممکن است رابطه‌ای متفاوت بین انتشار CO_2 و رشد اقتصادی داشته باشند. برای مثال، برخی از کشورهای درحال توسعه ممکن است منابع انرژی تجدیدپذیر بیشتری نسبت به کشورهای توسعه یافته داشته باشند و بنابراین سطح پایین‌تری از انتشار CO_2 را نیز خواهند داشت (استلیاروا، ۲۰۱۳، ص۳). همین استدلال برای کشورهای منطقه‌ای یکسان نیز صادق است. بنابراین به منظور انجام تحلیل داده‌های تابلویی از عوامل مؤثر بر انتشار گازهای گلخانه‌ای، علاوه بر مواردی که در بالا ذکر شد، چند عامل مؤثر دیگر باید در گروه‌بندی کشورها درنظرگرفته شود. مانند موهبت داشتن منابع سوخت‌های فسیلی، ترکیب انرژی[۶] (درصد استفاده از انرژی‌های جایگزین و هسته‌ای[۷])، اندازه کشور، سیاست‌های انرژی و محیط‌زیستی، سطح تولید و غیره (استلیاروا، ۲۰۱۳). براین اساس، در این مطالعه سعی بر این است تا به منظور درنظرگرفتن بیشترین معیارهای مؤثر بر ایجاد همگنی بین کشورها در مطالعات محیط‌زیستی از تحلیل خوشه‌ای[۸] که یکی از پرکاربردترین روش‌های آماری در طبقه‌بندی عناصر مختلف است بهره برده و به بررسی نتایج حاصل از آن پرداخته شود.

---

1. Organization of the Petroleum Exporting Countries
2. Belk
3. Organisation for Economic Co-Operation and Development
4. Pao & Tsai
5. Countries of Brazil, Russia, India and China
6. Energy Mix
7. Alternative and Nuclear Energy Use (ANEU)
8. Cluster Analysis

۲. پیشینه نظری و تجربی خوشه‌بندی در مطالعات اقتصادی

کپتنیئس[1] (۲۰۰۶) برای تشخیص مجموعه سری‌های قابل ادغام در مطالعات داده‌های تابلویی، روش معیار اطلاعات[2] را که برمبنای خوشه‌بندی سلسله‌مراتبی[3] قرار دارد، پیشنهاد می‌دهد. وی ویژگی‌های نظری این روش را موردبحث قرار داده و به بررسی کارایی نمونه کوچکی از آن، در مطالعه مونت کارلو[4] می‌پردازد.

وودریج[5] (۲۰۰۶) روش‌های خوشه‌بندی نمونه در اقتصادسنجی کاربردی را موردمطالعه قرار داده است. این مطالعه که مبسوطی از ویرایش ۲۰۰۳ آن است، به تفصیل جزئیات برای مدل‌های خطی پرداخته و خلاصه‌ای از استراتژی‌ها، (با تعداد خوشه‌های کم) در مدل‌های غیرخطی ارائه می‌دهد.

لیاو و هی[6] (۲۰۱۸)، در مطالعه خود برای افزایش کارائی انرژی و کاهش انتشار در زیربخش‌های صنایع مختلف در چین، با استناد به این مهم که زیربخش‌های صنایع همگن نیستند با خوشه‌بندی زیربخش‌های صنایع پرداخته و با بررسی تاثیر عوامل مختلف در کاهش انتشار به ارائه راهکار در این زمینه پرداخته‌اند.

سارافیدز و وبر[7] (۲۰۰۹) از روش خوشه‌بندی تقسیمی استفاده کردند و الگوریتم *تپه‌نوردی*[8] را که سعی می‌کند بهترین تفکیک براساس تعداد از پیش تعیین شده گروه را پیدا کند به کار بردند (استلیاروا،۲۰۱۳،ص.۵).

در مطالعات داخلی که از خوشه‌بندی داده‌های برای تحلیل استفاده شده است می‌توان به مطالعه رفعت (۱۳۹۷)، اشاره کرد که در مطالعه‌ای با عنوان «کاربرد تکنیک تحلیل شبکه در بررسی روابط چندجانبه تجاری ایران با مهم‌ترین شرکای تجاری در آسیا» بررسی الگوی تجاری ایران و شرکای عمده آن در آسیا پرداخته و بر مبنای معیارهای تعریف شده

1. Kapetanious
2. Information Criterion Method
3. Hierarchical Clustering
4. Monte Carlo
5. Woddridge
6. Nuo Liao and Yong He
7. Sarafidis & Weber
8. Hill-Climbing

به خوشه‌بندی طرف‌های تجاری اقدام نموده است. همچنین در یک مطالعه در حوزه بیمه متکی بر داده‌های پیمایش میدانی، مهدوی و ماجد (۱۳۹۰)، به خوشه‌بندی متقاضیان بیمه عمر و عوامل موثر بر آن پرداخته‌اند و تاثیرات غیرخطی سطوح مختلف مولفه‌های کمی بر تقاضای بیمه عمر را بیان کرده‌اند. مطالعات مربوط به خوشه‌بندی اشاره شده در تحقیقات اقتصادی داخل کشور، لزوماً در حوزه محیط‌زیست و انرژی نبوده و از روش‌های مختلف و نسبتاً ساده انجام خوشه‌بندی بهره جسته شده است.

یکی از مطالعات برجسته در زمینه بهره‌گیری در ورش خوشه‌بندی مطالعه استلیاروا (۲۰۱۳) است که رابطه علی پویا بین انتشار CO_2، رشد اقتصادی و درصد استفاده از انرژی‌های جایگزین و هسته‌ای (ANEU) را برای ۹۳ کشور منتخب طی دوره ۱۹۶۰-۲۰۰۸ است. وی از درصد استفاده از انرژی‌های جایگزین و هسته‌ای با عنوان ترکیب انرژی یاد کرده و روش خوشه‌بندی سلسله‌مراتبی برمبنای مؤلفه‌های اصلی را برای دسته‌بندی کشورها در گروه‌هایی همگن پیشنهاد می‌کند. استلیاروا برای رفع مشکل داده‌های پرت که نتایج خوشه‌بندی را به شدت تحت تأثیر قرار می‌دهد، متغیرها را طبقه‌بندی کرده و به منظور استخراج مؤلفه‌های اصلی از تحلیل تناظر چندمتغیره[1] (MCA) که برای تقلیل ابعاد (کدگذاری) متغیرهای طبقه‌بندی شده به کار می‌رود، استفاده می‌کند.

ایده خوشه‌بندی کشورها در اینجا از مطالعه و معیارهای مورداستفاده استلیاروا اخذ شده است، اما نحوه انجام خوشه‌بندی و نتایج آن متفاوت است. در این مطالعه ضمن معرفی مختصر تحلیل خوشه‌ای و تحلیل عاملی[2] نواقص مطالعه مذکور در زمینه خوشه‌بندی مرتفع و نتایج حاصل از آن بررسی شده است. جدول(۱) تفاوت‌های موجود در روش انجام خوشه‌بندی را در مطالعه حاضر با مطالعه استلیاروا به‌طورخلاصه نشان می‌دهد.

1. Multiple Correspondence Analysis (MCA)
2. Factor Analysis

جدول ۱. تفاوت‌های موجود در روش خوشه‌بندی مورداستفاده با مطالعه استلیاروا

مطالعه استلیاروا (۲۰۱۳)	مطالعه حاضر (۲۰۱٦)
• استفاده از داده‌های طبقه‌بندی شده برای خوشه‌بندی به منظور رفع مشکل داده‌های پرت	• استفاده از داده‌های رتبه‌بندی شده برای خوشه‌بندی به منظور رفع مشکل داده های پرت
• استفاده از داده‌های آخرین سال موردمطالعه (۲۰۰۸)	• استفاده از میانگین داده‌ها طی دوره ۲۰۱۲-۱۹۹۰
• استفاده از تابع MCA در تحلیل عاملی	• استفاده از تابع PCA در تحلیل عاملی*

*معایب و مزیت‌های هر مورد در صفحات ۱۰ تا ۱۲ توضیح داده‌شده است.

۲-۱. تحلیل خوشه‌ای (خوشه‌بندی)

تحلیل خوشه‌ای (که اولین بار توسط تراین[1] در سال ۱۹۳۹ مورد استفاده قرار گرفت) یکی از روش‌های آماری پرکاربرد در بسیاری از شاخه‌های علمی است. این روش به محقق امکان می‌دهد تا بر مبنای همگنی موجود در بین موارد یا موضوعات موردمطالعه، آنها را به شیوه‌ای مناسب، گروهبندی (طبقه‌بندی) نماید. به‌طوری‌که در این گروه‌بندی، موضوعات درون گروه شباهت زیادی با همدیگر داشته، اما تفاوت قابل‌توجهی با گروه‌های دیگر دارند.

۲-۱-۱. تمایز خوشه‌بندی از دسته‌بندی و کاربردهای آن

در دسته‌بندی براساس یک مدل، هر کدام از داده‌ها به دسته‌ای از پیش تعیین شده اختصاص می‌یابد. این دسته‌ها از طریق پژوهش‌های پیشین تعیین شده‌اند، لیکن در روش خوشه‌بندی هیچ دسته از پیش تعیین شده‌ای وجود ندارد و داده‌ها صرفاً براساس تشابه، گروه‌بندی می‌شوند و عناوین هر گروه نیز توسط محقق تعیین می‌شود. تکنیک‌های خوشه‌بندی به طور گسترده در مسائل تحقیقی متنوع به کار برده می‌شوند. هارتیگان[2] (۱۹۷۵) خلاصه‌ای از گزارش اکثر پژوهش‌های منتشرشده را به عنوان نتایجی از تحلیل‌های خوشه‌ای تهیه نمود. به‌طورکلی، هر زمان که شخص به خوشه‌بندی کوهی از اطلاعات در

1. Tryon
2. Hartigan

میان دسته‌های معنی‌دار قابل‌کنترل نیاز پیدا کند، تحلیل خوشه‌ای سودمند می‌باشد (صادق‌پور و مرادی،۱۳۸۶،ص.۱۷۴).

۲-۱-۲. نحوه اندازه‌گیری همگنی موارد و موضوعات در تحلیل خوشه‌ای

همگنی موضوعات و موارد می‌تواند به شیوه‌های مختلف تعیین شود. از روش‌های پرکاربرد برای اندازه‌گیری همگنی و شباهت، محاسبه فاصله اقلیدسی[1] است (کلانتری،۱۳۸۷،ص.۳۳۴). فاصله اقلیدسی در واقع فاصله هندسی در فضای چندبعدی است، که توسط فرمول (۱) محاسبه می‌شود (صادق‌پور و مرادی،۱۳۸۶،ص.۱۷۶):

$$\text{Distance(x,y)} = \{\sum_i (x_i - y_i)^2\}^{1/2} \qquad (۱)$$

باید توجه کرد، محدودیتی که در استفاده از فاصله اقلیدسی وجود دارد این است که، این روش به واحدهای اندازه‌گیری یا مقیاس متغیرها وابستگی شدید دارد. برای مثال وقتی متغیرها بر پایه مقیاس‌های مختلف (مانند قد و وزن) اندازه‌گیری می‌شوند، متغیرهایی که با اعداد بزرگتر اندازه‌گیری شده‌اند (مانند قد) نسبت به متغیرهایی با اعداد کوچکتر (مانند وزن)، سهم بیشتری در فاصله اندازه خواهند داشت. البته برای رفع این محدودیت، محقق می‌تواند داده‌های خام را با استفاده از روش‌های مختلف استانداردسازی، استاندارد کرده تا از این طریق به رفع اختلاف مقیاس متغیرها اقدام کند و سپس به محاسبه فاصله اقلیدسی بپردازد (کلانتری،۱۳۸۷،ص.۳۳۵).

۲-۱-۳. تکنیک‌های خوشه بندی

پس از اتخاذ تصمیم بر استفاده از معیار فاصله و محاسبه فواصل براساس آن، قدم بعدی نحوه طبقه‌بندی موضوعات براساس این فواصل است که به آن تکنیک خوشه‌بندی می‌گویند. تکنیک‌های عمومی مورد استفاده در تحلیل خوشه‌ای به دو دسته: الف)خوشه‌بندی سلسله‌مراتبی و ب) خوشه‌بندی تقسیمی یا غیرسلسله‌مراتبی[2] تقسیم می‌شوند.

1. Euclidean Distance
2. Partitioning or Nonherarchical Clustering

الف) خوشه‌بندی سلسله مراتبی: خوشه‌بندی سلسله مراتبی یک ساختار سلسله مراتبی از خوشه‌ها ایجاد می کند، به بیان دیگر، یک درخت از خوشه‌ها که به درختواره‌نگار[1] معروف است. هر گره یک خوشه است که خود شامل خوشه های فرزند است. روش‌های خوشه‌بندی سلسله مراتبی به دو دسته: ۱) تراکمی یا پایین به بالا[2] و ۲) تفکیک‌پذیری یا بالا به پایین[3]، قابل تقسیم است. در خوشه‌بندی تراکمی کار را با خوشه‌هایی متشکل از یک داده یا متغیر شروع می‌کنیم و بطور تکراری دو یا چند خوشه متناسب را با هم تلفیق می‌کنیم. بسیاری از برنامه‌های رایانه‌ای روش تراکمی را انجام می‌دهند. در روش تفکیک‌پذیری عکس روش تراکمی ابتدا همه افراد در یک خوشه منظور می‌شوند و سپس مرحله به مرحله خوشه‌های جدید ایجاد می‌شود. چهار روش عمده که در نحوه ادغام خوشه‌ها در هم یا نحوه محاسبه فاصله بین خوشه‌ها با یکدیگر متفاوتند، برای تشکیل خوشه‌های تراکمی مورداستفاده قرار می‌گیرد. این روش‌ها که قواعد پیوند یا ادغام نامیده می‌شوند، عبارتند از:

۱. پیوند تکی یا نزدیکترین همسایه[4]
۲. پیوند کامل یا دورترین همسایه[5]
۳. پیوند متوسط[6]
۴. Ward: این روش از سه روش قبلی ممتاز و کاراتر است، زیرا از آنالیز واریانس برای سنجش فواصل میان خوشه‌ها استفاده می‌نماید. به طور مختصر این روش سعی می‌کند مجموع مربعات را برای هر دو خوشه‌ای که می‌تواند در هر گام تشکیل دهد، به حداقل برساند (صادق‌پور و مرادی،۱۳۸۶،ص.۱۸۰). به همین دلیل آن را روش کمترین واریانس۷ نیز می‌گویند.

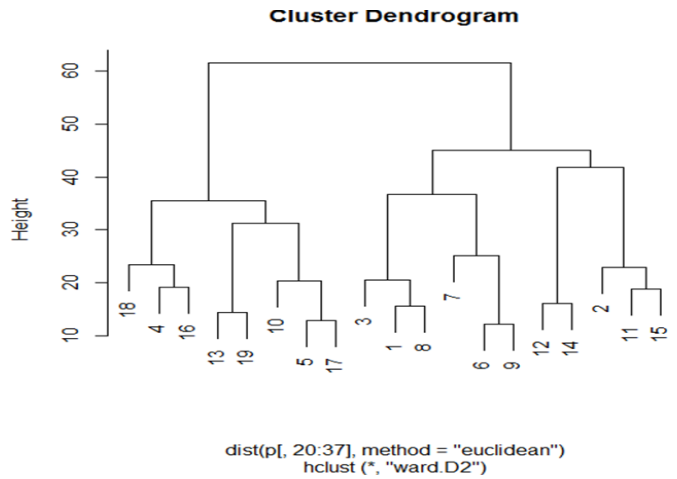
1. Dendrogram
2. Agglomerative or Bottom-Up
3. Divisive or Top-Down
4. Single Linkage or Nearest Neighbour
5. Complete Linkage or Furthest Neighbour
6. Average Linkage
7. Minimum Variance

نموداری که مراحل تشکیل خوشه‌ها را به صورت سلسله‌مراتبی نشان می‌دهد و به عنوان یک ابزار مهم در تعیین تعداد مناسب خوشه‌ها مورد استفاده قرار می‌گیرد، نمودار درختی یا درختواره‌نگار نامیده می‌شود. نمودار(۱)، نمودار درختی مربوط به خوشه‌بندی سلسله مراتبی با فاصله اقلیدسی و روش ادغام Ward را نشان می‌دهد.

ب) خوشه‌بندی تقسیمی یا غیرسلسله مراتبی: در مقابل روش‌های سلسله‌مراتبی، فرآیند خوشه‌بندی غیرسلسله مراتبی یا تقسیمی قرار دارد که از یک ساختار درختی تبعیت نمی‌کند، بلکه در این روش اولین قدم، انتخاب مبدأ و بذر اولیه است. با تعیین این بذر اولیه، کلیه افراد درون یک فاصله آستانه‌ای در آن خوشه قرار می‌گیرند (کلانتری،۱۳۸۷،ص.۳۴۲). بنابراین در این تکنیک‌ها، یا تعداد خوشه‌ها از قبل مشخص است و یا اینکه باید از روی داده‌ها مشخص شود. یکی از معروف‌ترین این تکنیک‌ها، روش K-میانگین[1] نامیده می‌شود. علت اطلاق نام K-میانگین به این روش آن است که هر فرد به خوشه‌ای منتسب می‌شود که به میانگین آن خوشه نزدیک‌تر است و دقیقاً K خوشه مختلف که بیشترین تفاوت را با یکدیگر دارند ارائه می‌دهد.

در صورتی که تعداد افراد خیلی زیاد باشد از الگوریتم K -میانگین و یا hill-Climbing (این الگوریتم از لحاظ مراحل شبیه الگوریتم K-میانگین است اما معیارهای این دو فرق می‌کند) استفاده می‌شود.

1.K-means



نمودار۱. نمودار درختی سلسله مراتبی با فاصله اقلیدسی و روش ادغام Ward

۲-۲. تحلیل عاملی

تحلیل عاملی نامی عمومی است برای برخی از روش‌های آماری چندمتغیره که هدف اصلی آن خلاصه کردن داده‌ها در تعداد محدودی عامل است، به طوریکه در این فرآیند کمترین میزان گم شدن اطلاعات وجود داشته‌باشد (کلانتری،۱۳۸۷،ص.۲۸۴).

در تحلیل عاملی مدل‌های مختلفی وجود دارد که از میان آن‌ها دو روش تحلیل مؤلفه‌های اصلی[1] (PCA) و تحلیل عاملی مشترک[2] از پرکاربردترین این روش‌ها هستند. انتخاب هر یک از این مدل‌ها به هدف محقق بستگی دارد. زمانیکه محقق درصدد پیش‌بینی و تعیین کمترین تعداد عامل‌هاست که قادر باشد بیشترین واریانس موجود در مقادیر اصلی را تبیین کند، روش تحلیل مؤلفه‌های اصلی استفاده می‌شود. در مقابل تحلیل عاملی مشترک، زمانی به کار می‌رود که هدف، شناسایی عامل یا ابعادی باشد که به سادگی قابل شناسایی نیستند (کلانتری،۱۳۸۷،ص.۲۹۵).

تحلیل مؤلفه‌های اصلی در سال ۱۹۰۱ توسط پیرسون[3] ارائه شده است. این تحلیل شامل تجزیه مقادیر ویژه ماتریس کوواریانس است که به منظور کاهش ابعاد مجموعه داده‌ها مورد

1. Principal Component Analysis (PCA)
2. Common Factor Analysis
3. Pearson

استفاده قرار می‌گیرد. به عبارت دیگر تحلیل مؤلفه‌های اصلی می‌تواند به عنوان یک روش رفع اخلال[1] که اخلال[2] را از داده‌ها جدا می‌کند، استفاده شود. همچنین برای نمایش درخت سلسله مراتبی (درختواره‌نگار) در یک نقشه سه بعدی که به درک بهتر (موقعیت) داده‌ها (در هر خوشه) کمک می‌کند (هاسن، جز و پجز[3]،۲۰۱۰،ص.۱).

تعداد مؤلفه‌ها یا ابعاد استخراج شده در هر مدل برابر است با تعداد معیارهایی (متغیرهایی) که بررسی می‌شوند. اولین ابعاد، عمده اطلاعات (اطلاعات ضروری) را استخراج می‌کند و سهم بزرگی از کل واریانس را به خود اختصاص می‌دهند، در حالی که آخرین ابعاد از کمترین میزان واریانس برخوردار بوده و به اخلال محدود می‌شوند. بنابراین با حذف اخلال از داده‌ها، خوشه‌بندی بر مبنای مؤلفه‌های اصلی از دقت و استحکام (اعتبار) بیشتری نسبت به خوشه‌بندی بدست آمده از مقادیر اصلی داده‌ها برخوردار است (هاسن، جز و پجز، ۲۰۱۰، ص.۲).

تعداد ابعاد باقی‌مانده برای خوشه‌بندی، می‌تواند با روش‌های متعددی انتخاب شود (جولایف، ۲۰۰۲) که مهمترین آنها عبارتند از:

الف. معیار مقدار ویژه[4]: براساس این معیار تنها، عامل‌هایی که مقدار ویژه آنها بیشتر از یک باشد به عنوان عامل‌های معنی‌دار درنظرگرفته می‌شوند و کلیه عامل‌هایی که مقدار ویژه آنها کمتر از یک است به عنوان عامل‌هایی که از نظر آماری معنی‌دار نیستند از تحلیل کنار گذاشته می‌شوند (کلانتری، ۱۳۸۷، ص.۳۰۴). بسیاری از محققان از جمله کیسر[5] مقدار ویژه‌ی یک را مبنای تعیین تعداد عامل‌ها قرار می‌دهند.

ب. معیار درصد واریانس: در این رهیافت درصد تجمعی واریانس، مبنای تصمیم‌گیری قرار می‌گیرد. در تحقیقات غیرعلوم انسانی، فرآیند استخراج عوامل باید تا زمانی ادامه یابد که این عامل‌ها بتوانند حداقل ۹۵ درصد واریانس داده‌ها را تبیین نمایند. اما در تحقیقات اجتماعی-انسانی تبیین ۶۰ درصد واریانس نیز کفایت می‌کند.

1. Denoising
2. Noise
3. Husson, Josse & Pagès
4. Eigenvalue Criterion
5. Kaiser

۳. روش‌شناسی پژوهش، نمونه موردبررسی و معیارهای خوشه‌بندی

در این پژوهش با توجه به هدف مطالعه و حجم نمونه موردبررسی از تکنیک سلسله‌مراتبی استفاده شده و به منظور استحکام نتایج، خوشه‌بندی براساس مؤلفه‌های اصلی بدست‌آمده انجام گرفته است. درجه ناهمانندی را به کمک فاصله اقلیدسی اندازه‌گیری کرده و از روش ادغام Ward برای ترکیب خوشه‌ها بهره بردیم. جهت انجام تحلیل عاملی و تحلیل خوشه‌ای نیز از نرم‌افزار R 3.3.0 استفاده نمودیم.

توابع PCA و HCPC واقع در بسته FactoMineR (بسته‌ای که به تحلیل اکتشافی داده‌ها در نرم‌افزار R اختصاص دارد) به ترتیب برای استخراج مؤلفه‌های اصلی و جهت انجام خوشه‌بندی سلسله‌مراتبی بر مبنای مؤلفه‌های اصلی، مورداستفاده قرار گرفتند.

نمونه موردبررسی ۹۲ کشور با بیشترین میزان انتشار CO_2 است. معیارهای تحلیل عاملی و تحلیل خوشه‌ای نیز که از میانگین آن‌ها طی دوره ۲۰۱۲-۱۹۹۰ استفاده شده، در جدول(۲) ارائه شده‌است (از نظر آماری میانگین داده‌ها نسبت به داده‌های یک سال موردنظر، معرف بهتری برای نشان دادن ویژگی‌های نمونه است. زیرا به عنوان مثال در این مورد خاص ممکن است میزان انتشارCO_2 ، تولید ناخالص داخلی یا هرکدام از معیارهای دیگر در یک سال موردنظر به دلایلی نظیر رکود یا رونق اقتصادی، عوامل سیاسی، تغییرات آب‌وهوایی و غیره دارای تورش به سمت بالا یا پایین باشند، بنابراین به نظر می‌رسد استفاده از میانگین داده‌ها یکی از مزیت‌های این مطالعه نسبت به مطالعه استلیاروا است). این معیارها، علاوه بر انتشار CO_2 (کل و سرانه)، رشد اقتصادی (GDP کل و سرانه) و منطقه، سایر عوامل مؤثر بر ایجاد همگنی بین کشورها از جمله موهبت داشتن منابع سوخت‌های فسیلی (نفت، گاز و ذغال‌سنگ)، ترکیب انرژی (درصد استفاده از انرژی‌های جایگزین و هسته‌ای)، درصد تولید برق از منابع متفاوت (درصد تولید برق از نفت، درصد تولید برق از گاز و ...)، جمعیت و غیره که جمعاً ۱۸ معیار بوده و برگرفته از مقاله استلیاروا (۲۰۱۳) است را دربرمی‌گیرد با این تفاوت که همان‌طورکه در بالا ذکر شد از میانگین داده‌ها استفاده شده است.

لازم به ذکر است که در میان کشورهای موردمطالعه، چند کشور داده‌های پرت دارند (بزرگترین کشورها مانند چین، آمریکا، روسیه و کانادا که سطح بالایی از انتشار CO_2،

درآمد و جمعیت را دارند و مقدار زیادی سوخت فسیلی تولید می‌کنند) که نادیده گرفتن آنها ممکن نیست و نمی‌توانند بدون تحریف نتایج خوشه‌بندی موردمطالعه قرار گیرند.

جدول ۲. معیارهای مورد استفاده در خوشه‌بندی

متغیر	منبع	توضیح	واحد
CO_2	WDI[1]	انتشار CO_2	متریک تن به نفر (Mt)
KCO_2	WDI	انتشار سرانه CO_2	متریک تن به نفر (Mt)
GDP	WDI	تولید ناخالص داخلی	میلیون دلار به قیمت‌های ثابت ۲۰۱۰
KGDP	WDI	تولید ناخالص داخلی سرانه	میلیون دلار به قیمت‌های ثابت ۲۰۱۰
POP	WDI	جمعیت	میلیون نفر
ANEU	WDI	استفاده از انرژی جایگزین و هسته‌ای	به درصد کل مصرف انرژی
CO_2-elec	WDI	انتشار CO_2 از بخش نیروگاهی	به درصد کل انتشار CO_2
CO_2-indus	WDI	انتشار CO_2 از بخش صنعتی	به درصد کل انتشار CO_2
Elec-coal	WDI	تولید برق از ذغال‌سنگ	به درصد کل تولید برق
Elec-oil	WDI	تولید برق از نفت	به درصد کل تولید برق
Elec-hydro	WDI	تولید برق از انرژی برق آبی	به درصد کل تولید برق
Elec-gas	WDI	تولید برق از گاز	به درصد کل تولید برق
Elec-nucl	WDI	تولید برق از انرژی هسته‌ای	به درصد کل تولید برق
Elec-RES	WDI	تولید برق از منابع انرژی تجدیدپذیر به‌جز انرژی برق آبی	به درصد کل تولید برق
Prod-oil	BP[2]	تولید نفت	میلیون تن (Mt)
Prod-gas	BP	تولید گازطبیعی	میلیون تن (Mt)
Prod-coal	BP	تولید ذغال‌سنگ	میلیون تن معادل با نفت (Mteo)
Regoin	WDI	منطقه	آسیای شرقی و اقیانوسیه، آسیای جنوبی، آمریکای لاتین و کارائیب، آمریکای شمالی، اروپا وآسیای مرکزی، خاورمیانه و شمال آفریقا، صحرای آفریقا

منبع: استلیاروا، (۲۰۱۳)

1. World Development Indicators (WDI)
2. British Petroleum

استلیاروا (۲۰۱۳) برای رفع این مشکل متغیرها را به چهار طبقه خیلی‌پایین، پایین، بالا و خیلی بالا تقسیم‌بندی نموده‌است و در تحلیل عاملی به جای استفاده از تابع PCA که برای متغیرهای کمی به کار می‌رود، از تابع MCA که مختص متغیرهای طبقه‌بندی شده است، استفاده نموده است. اما به نظر می‌رسد این راه‌حل چندان مناسب نباشد، زیرا افرادی را که در یک طبقه قرار دارند از نظر برخورداری از آن ویژگی خاص یکسان در نظر می‌گیرد. برای مثال کشورهای چین، آمریکا، برزیل، ایران، تایلند و چند کشور دیگر را از نظر انتشار CO_2 در طبقه خیلی بالا قرار داده و سایر کشورها نیز در طبقات دیگر تقسیم شده‌اند. با این کار کشورهایی مانند چین و تایلند که در میانگین انتشار CO_2، ۶۰۸۰ واحد (متریک تن) با یکدیگر تفاوت دارند (چین با میانگین انتشار ۶۳۰۴ واحد و تایلند ۲۲۴ واحد) یکسان در نظر گرفته شده و به فاصله زیاد آنها با یکدیگر هیچ توجهی نمی‌شود، که این مسئله می‌تواند به عدم دستیابی محقق به مطلوب‌ترین خوشه‌بندی ممکن، منجر شود. بنابراین برای استاندارد کردن متغیرها روش رتبه‌بندی که در فرمول (۲) ارائه شده به کار برده شده است که علاوه بر یکسان نمودن مقیاس متغیرها، مشکل فوق را نیز برطرف می‌نماید و دیگر احتیاجی به استفاده از تابع MCA به جای تابع PCA نیست. در این روش فرد iام براساس میزان برخورداری از صفت jام، در بین تمام افراد رتبه‌بندی می‌شود. برای مثال ۹۲ کشور براساس میزان انتشار CO_2، از کشور با رتبه یک که دارای بیشترین میزان انتشار است تا کشور با رتبه ۹۲ که دارای کمترین میزان انتشار است، رتبه‌بندی می‌شوند.

$$X_{ij}^{*} = Rank_i(X_{ij}) \qquad (۲)$$

٤. یافته‌های پژوهش

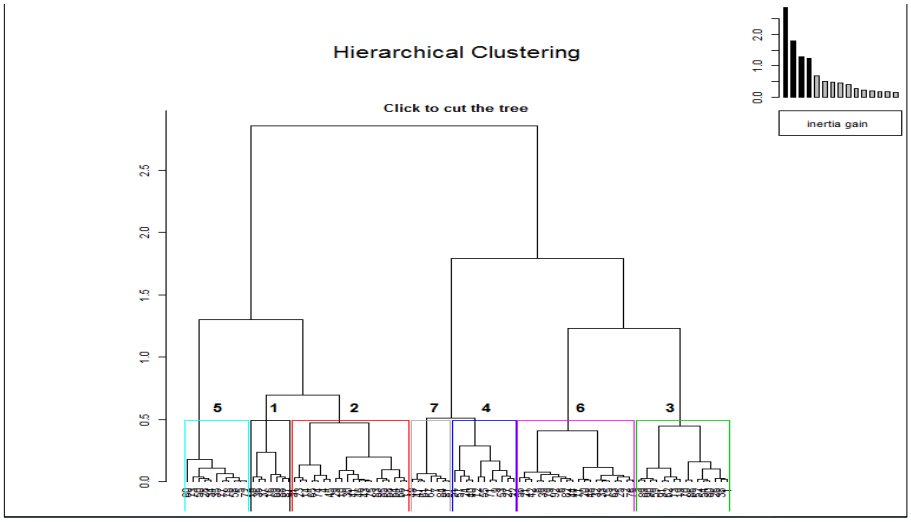
۱۸ معیار معرفی شده برای انجام خوشه‌بندی، با استفاده از تحلیل عاملی و براساس معیار مقدار ویژه به پنج عامل تقلیل یافتند که جمعاً ۷۵ درصد از واریانس را در بر می‌گیرند. جدول(۳) نتایج حاصل از تحلیل عاملی را ارائه می‌دهد. پس از استخراج عامل‌ها، خوشه‌بندی با استفاده از تابع HCPC براساس پنج عامل بدست آمده انجام شده‌است. این

تابع، تعداد بهینه خوشه را براساس افزایش واریانس[1] بین گروهی پیشنهاد می‌نماید، که با نگاه کردن به نمودار درختی نیز مشخص می‌شود. اما انتخاب تعداد خوشه‌ها با توجه به اهداف پژوهش و نظر محقق به عهده وی است، بنابراین با درنظرگرفتن بهترین ترکیب برای هر گروه، درخت سلسله مراتبی با هفت خوشه قطع شده است. نمودارهای (۲) و (۳) نتیجه خوشه‌بندی را به ترتیب در نمودار درختی دوبعدی و سه‌بعدی نشان می‌دهند. ترکیب هر خوشه با ویژگی‌های آن نیز در جدول ۳ ارائه شده‌است.

جدول ۳. نتایج تحلیل عاملی

فهرست عامل‌ها	مقدار ویژه	درصد واریانس	درصد واریانس تجمعی
۱	۴/۳۵۰	۲۴/۱۶۴	۲۴/۱۶۴
۲	۳/۴۸۴	۱۹/۳۵۳	۴۳/۵۱۷
۳	۲/۹۸۶	۱۶/۵۹۱	۶۰/۱۰۸
٤	۱/۶۶۶	۹/۲۵۳	۶۹/۳۶۱
۵	۱/۰۴۳	۵/۷۹۴	۷۵/۱۵۵

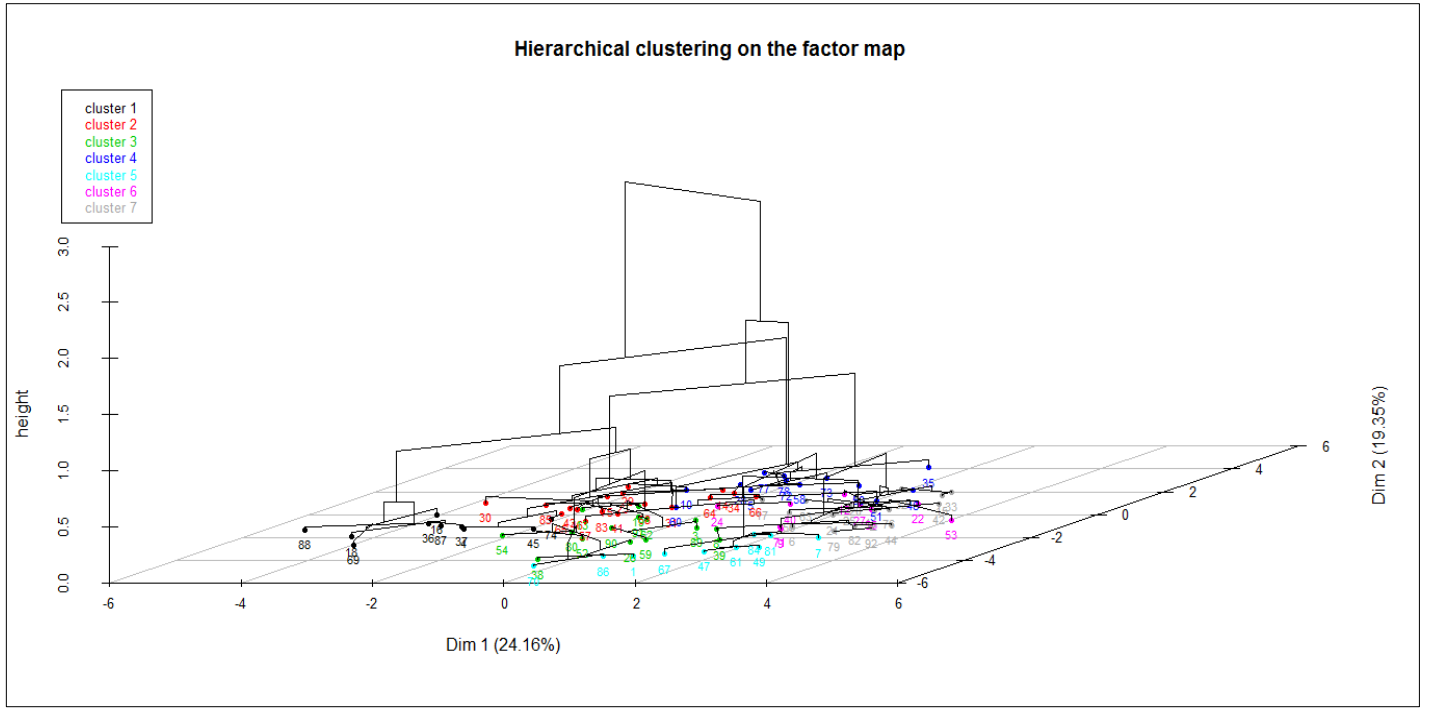
منبع: یافته‌های پژوهش



نمودار۲. نمودار درختی دوبعدی حاصل از خوشه‌بندی با تابع HCPC

منبع: یافته‌های پژوهش بر مبنای خروجی نرم‌افزار R

1. Inertia gain



نمودار۳. نمودار درختی سه‌بعدی حاصل از خوشه‌بندی به روش HCPC

منبع: یافته‌های پژوهش بر مبنای خروجی نرم‌افزار R

جدول ٤. ترکیب خوشه‌ها و مشخصات آن‌ها

خوشه	ترکیب خوشه	مشخصات خوشه (برای اکثر کشورها)	کشورنماینده
۱	آفریقای جنوبی، آمریکا، استرالیا، اندونزی، انگلستان، چین، روسیه، قزاقزستان، کانادا، هند	- بزرگترین کشورها (از نظر اقتصاد، مساحت، جمعیت، تولید برق و ...) - با سطح **بسیار** بالای انتشار CO_2 و GDP - بزرگترین تولیدکنندگان سوخت‌های فسیلی (مجموع نفت، گاز و ذغال‌سنگ)	**آمریکا**
۲	آلمان، اکراین، اسپانیا، ایتالیا، بلغارستان، پرتغال، ترکیه، چک، رومانی، ژاپن، فرانسه، فیلیپین، کره جنوبی، لهستان، مجارستان، هلند، یونان	- کشورهایی از اروپای شرقی - با سطح بالای انتشار سرانه CO_2 و GDP سرانه - درصد بالای تولید برق از ذغال سنگ و تولیدکنندگان عمده برق هسته‌ای - درصد بالای ANEU	**اکراین**
۳	آرژانتین، ازبکستان، ایران، برزیل، بنگلادش، پاکستان، تایلند، عراق، کلمبیا، مالزی، مصر، مکزیک، نیجریه، ونزوئلا، ویتنام	- تولیدکنندگان سوخت های فسیلی بویژه گازطبیعی - با درصد بالای تولید برق از گازطبیعی - سطح پایین انتشار سرانه CO_2 (اما انتشار CO_2 بالا)	**مکزیک**
۴	اتریش، اسلواکی، اسلونی، ایسلند، بلژیک، سوئد، سوئیس، فنلاند، لاتویا، لوکزامبورگ، لیتوانی، نروژ، نیوزلند	- کشورهایی از اروپای غربی - با سطح **بسیار** بالای GDP سرانه و انتشار بالای سرانه CO_2 (اما انتشار CO_2 پایین) - درصد **بسیار** بالای ANEU	**سوئیس**
۵	الجزایر، امارات، بحرین، ترکمنستان، ترینداد و توباگو، عربستان، عمان، کویت، قطر، لیبی	- کشورهایی از خاورمیانه - با سطح **بسیار** بالای انتشار سرانه CO_2 و GDP سرانه - تولیدکنندگان عمده و صادرکننده نفت (عضو اوپک) - بدون ANEU	**قطر**

خوشه	ترکیب خوشه	مشخصات خوشه (برای اکثر کشورها)	کشورنماینده
۶	استونی، ایرلند، بلاروس، بوسنی و هرزگوین، دانمارک، سنگاپور، قبرس، مالتا، مغولستان	- کشورهایی کوچک و جزیره ای - با سطح بالای انتشار سرانه CO_2 - درصد بسیار پایین ANEU	**استونی**
۷	آذربایجان، آنگولا، اردن، اکوادور، بلیوی، پرو، تونس، جامائیکا، سودان، سوریه، شیلی، کامرون، کرواسی، کوبا، گوآتمالا، مراکش، هندوراس، یمن	- کشورهایی از آمریکای جنوبی و آفریقا - با سطح **بسیار** پایین انتشار سرانه CO_2 و GDP سرانه - درصد بالای تولید برق آبی	**سودان**

منبع: یافته‌های تحقیق

۵. نتیجه‌گیری

در مطالعات چندمقطعی، خوشه‌بندی و یا دسته‌بندی مناسب و صحیح افراد، بنگاه‌ها و کشورها در گروه‌هایی همگن، به منظور دارا بودن اعتبار لازم از لحاظ آماری دارای اهمیت بسیاری است. پژوهش حاضر به دنبال ارائه روشی برای خوشه‌بندی کشورها در مطالعات بین‌کشوری بویژه در مطالعات اقتصادسنجی مبتنی بر روش داده‌های تابلویی است. با توجه به اهداف مختلف مطالعات اقتصادسنجی، گروه‌بندی کشورها برمبنای مؤلفه‌ها و شاخص‌های تاثیرگذار و مد نظر پژوهش انجام می‌پذیرد. این پژوهش نیز براساس مؤلفه‌های اقتصادی و محیط‌زیستی به علاوه سایر مؤلفه‌های مؤثر به خوشه‌بندی کشورها پرداخته است. خوشه‌بندی انجام شده بر مبنای اطلاعات در دسترس در بازه زمانی ۱۹۹۰ تا ۲۰۱۲ صورت پذیرفته. ۹۲ کشور موردمطالعه به هفت خوشه متفاوت، هرکدام با ویژگی‌های خاص خودشان تفکیک شدند. نتایج خوشه‌بندی با بررسی مشخصات غالب کشورها در هر خوشه، نشان از همگنی موردنظر در هریک از خوشه‌های مشخص‌شده دارد، به‌عنوان مثال خوشه پنج متشکل از کشورهای عمده تولیدکننده و صادرکننده نفت در منطقه خاورمیانه، با سطح بسیار بالای انتشار سرانه CO_2، درآمد سرانه بالا و بدون استفاده از انرژی جایگزین و هسته‌ای است که تولید برق آن نیز تنها از منابع نفت و گاز صورت می‌گیرد. یا خوشه‌ی چهار که دربردارنده کشورهای اروپای غربی، با سطح بسیار بالای درآمد سرانه، درصد بسیار بالای استفاده از انرژی‌های جایگزین وهسته‌ای و سطح پایین انتشار CO_2 است. به این‌ترتیب می‌توان با توجه به هدف مطالعه، از هریک از خوشه‌های بدست آمده برای استفاده در تحلیل‌های مبتنی بر داده‌های تابلویی بهره برد و اطمینان داشت که گروه مورمطالعه همگن‌ترین اعضای ممکن را از لحاظ بیشترین معیارهای مؤثر در همگنی داراست. یا با مطالعه همزمان همه خوشه‌ها (با توجه به مشخصات بدست آمده برای هریک) به مقایسه نتایج حاصل از برآوردها در درون خوشه‌های مختلف با یکدیگر پرداخت. براین اساس، پیشنهاد می شود در سایر حوزه‌های مطالعات اقتصادی ازجمله انرژی، علوم مالی، توسعه، کشاورزی و ... نیز به‌منظور طبقه‌بندی مقاطع یا انتخاب

مناسب‌ترین گروه موردمطالعه برای استفاده در تحلیل‌های چند مقطعی بویژه تحلیل‌های مبتنی بر داده‌های تابلویی (حتی در مطالعات خارج از حوزه اقتصاد) از روش‌های آماری مانند انواع روش‌های خوشه‌بندی بهره برده تا بدین وسیله ضمن افزایش اطمینان از همگنی موردنظر و استحکام و اعتبار نتایج بدست‌آمده، امکان مقایسه گروه‌های مختلف با ویژگی‌های متفاوت نیز فراهم گردد.

منابع و مآخذ

- Bakirtas, I., Bayrak, S., & Cetin, A. (2014). Economic growth and carbon emission: A dynamic panel data analysis. *European Journal of Sustainable Development*, 3(4), 91-102.
- Belke, A., Dreger, Ch., Haan, F., (2010), Energy Consumption and Economic Growth –New Insights into the Cointegration Relationship, *Ruhr Economic Papers*, (190).
- Dinda. S., & Coondo. D. (2006), Income and emission: a panel data-based cointagration analysis. *Ecological Economics*, (57), 167-181.
- Du, K. (2018). Econometric convergence test and club clustering using Stata. *The Stata Journal*, 17(4), 882-900.
- ESSO, J.L. (2010), The Energy Consumption-Growth Nexus in Seven Sub-Saharan African Countries,, *Issue 2*,(30), 1191-1209.
- Huang, B., Hwang, M.J., Yang, C.W., (2008), Causal relationship between energy consumption and GDP growth revisited: A dynamic panel data approach, *Ecological Economics*, (67), 41-54.
- Husson, F., Josse, J., & Pagès. J. (2010). Principal component methods-hierarchical clustering-partitional clustering: why would we need to choose for visualizing data?, Agrocampus, 1.
- Kalantari, K. 2003. *Data Processing and Analysis in Socio-Economic Research*. Sharif Publication, Tehran. {In Persian}
- Kapetanios, G. (2005). Cluster Analysis of Panel Datasets using Non-Standard Optimisation of Information Criteria. Queen Mary, University of London, (535).
- Lee, C. & Chang, C. (2007), Energy consumption and economic growth in Asian countries: A more comprehensive analysis using panel data, *Resource and Energy Economics*.
- Liao, N., & He, Y. (2018). Exploring the effects of influencing factors on energy efficiency in industrial sector using cluster analysis and panel regression model. *Energy*, 158, 782-795.
- Mahdavi G, Majed V.(2011). The impact of Socio-Economic and Psychological Factors on Life Insurance Demand in Iran. *JEMR*. 2 (5) :21-46. .[Persian]
- McNeish, D., & Kelley, K. (2018). Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making recommendations. *Psychological methods*.
- Pao. H. T., & Tsai. C. M. (2010), "CO_2 emissions, energy consumption and economic growth in BRIC countries". *Energy Policy*, (38), 7850-7860.

- Pourkazemi, M., & Ebrahimi, I. (2008). Evaluation of environmental Kuznets curve in the Middle East. *Journal of Economic Research*, 34, 57-71. {In Persian}
- Rafat M. (2019). The Application of Complex Networks Analysis to Assess Iran's Trade and It's Most Important Trading Partners in Asia. *JEMR*. 9 (34) :107-137 {In Persian}.
- Squalli, J. (2006), Electricity consumption and economic growth: Bounds and causality analyses of OPEC members, *Energy economics*.
- Stolyarova, E. (2013). "Carbon Dioxide Emissions, economic growth andenergy mix: empirical evidence from 93 countries". *Climate Economics Chair*.
- Wolde-Rufael, Y. (2006), Electricity consumption and economic growth: a time series experience for 17 African countries, *Energy Policy*, (34), 1106-1114.
- Wooldrige, M. (2006). CLUSTER-SAMPLE METHODS IN APPLIED ECONOMETRICS: AN EXTENDED ANALYSIS. *Michigan State University*. (517) 353-5972.
- www.bp.com. Statistical Review of World Energy (2016)
- www.eia.gov. (2016)
- www.worldbank.org. World Development Indicator (2016)

Using Clustering and Factor Analysis in Cross Section Analysis Based on Economic-Environment Factors

Vahid Majed [1], Hossein Mirshojaeian Hosseini [2], Samira Riazi[3]

Received: 2018/06/4 Accepted: 2019/03/9

Abstract

Homogeneity of groups in studies those use cross section and multi-level data is important. Most studies in economics especially panel data analysis need some kinds of homogeneity to ensure validity of results. This paper represents the methods known as clustering and homogenization of groups in cross section studies based on enviro-economics components. For this, a sample of 92 countries which produce the most greenhouse gases including CO_2, clustered based on 18 criteria. Those criteria reduced to five primary components using factor analysis. Clustering of countries done by HCPC (Hierarchical Clustering on Principal Component) method. All 92 countries were clustered in 7 different groups. For each group properties of countries indicates the homogeneity of each cluster. In cross section analysis with many sections, especially analysis based on panel data, clustering, increases assurance of expected homogeneity and validity of result.

Keywords: Dentist Cluster Analysis, Factor Analysis, Cross Section Studies, Enviro-Economic Components.

JEL Classification: C23, Q56, O57.

1. Assistant Professor of Economics, University of Tehran, (Corresponding Author), Email:Majed@ut.ac.ir
2. Assistant Professor of Economics, Economic Affairs Research Institute, Email:hmirshojaeian@gmail.com
3. MA, Faculty of Economics, University of Tehran, Email:s.riazi@ut.ac.ir