فصلنامه پژوهش‌های اقتصادی ایران/ سال بیست‌ودوم/ شماره ۷۲/ پاییز ۱۳۹۶/ صفحات ۷۸-۴۱

شناسایی بخش‌های کلیدی اقتصاد ایران با استفاده از خوشه‌بندی فازی

اسفندیار جهانگرد[۱]

علیرضا ناصری بروچنی[۲]

تاریخ ارسال: ۱۳۹۵/۹/۱۱ تاریخ پذیرش: ۱۳۹۶/۸/۱

چکیده

در این مقاله با تلفیق الگوی داده- ستانده و خوشه‌بندی فازی، با ترکیب شاخص‌های بین‌بخشی، وزن اقتصادی و پتانسیل‌های اقتصادی هر بخش، بخش‌های کلیدی اقتصاد ایران با استفاده از جداول داده- ستانده سال‌های ۱۳۹۰ و ۱۳۸۵ شناسایی شده است. پیش از آن، برای حذف اثر نامطلوب داده‌های پرت بر خوشه‌بندی، داده‌های پرت شناسایی و به‌طور جداگانه خوشه‌بندی شده‌اند. نتایج به‌دست‌آمده نشان می‌دهد که بخش‌های کلیدی اقتصاد ایران در گروه داده‌های پرت قرار دارد و می‌توان از جداسازی این داده‌ها برای شناسایی بخش‌های کلیدی اقتصاد در این روش استفاده کرد. با تفسیر نتایج خوشه‌بندی و تحلیل شاخص‌های تعریف شده، بخش‌های صنعتی و معدنی (ساخت سایر محصولات کانی غیر فلزی، ساخت فلزات اساسی، ساخت ماشین‌آلات و تجهیزات طبقه‌بندی نشده در جای دیگر، توزیع گاز طبیعی، سایر معادن و ساخت کک، فرآورده‌های حاصل از تصفیه نفت و سوخت‌های هسته‌ای) بخش‌های کلیدی اقتصاد ایران هستند. همچنین بخش «نفت خام و گاز طبیعی» از نظر صادرات و بخش‌های خدمات (عمده‌فروشی، خرده‌فروشی، تعمیر وسایل نقلیه و کالاها، امور عمومی، حمل‌ونقل جاده‌ای) و «کشاورزی و باغداری» از نظر اشتغال‌زایی، بخش‌های کلیدی اقتصاد ایران شناسایی شدند.

واژگان کلیدی: بخش‌های کلیدی، خوشه‌بندی، منطق فازی، الگوریتم خوشه‌بندی فازی c-means نااقلیدسی (NERFCM)،تحلیل داده- ستانده

طبقه‌بندی JEL : *C38,D57*

۱- دانشیار گروه اقتصاد نظری دانشگاه علامه طباطبائی(نویسنده مسئول)، پست الکترونیکی:
ejahangard@gmail.com

۲- کارشناسی ارشد اقتصاد دانشگاه علامه طباطبائی، پست الکترنیکی:
alireza_nasseri2009@yahoo.com

۱- مقدمه

شناسایی بخش‌های کلیدی اقتصاد از منظر برنامه‌ریزی و تخصیص بهینه منابع محدود برای رسیدن به رشد پایدار و مناسب اهمیت دارد. بر مبنای تئوری‌های رشد نامتوازن و قطب رشد، سرمایه‌گذاری‌ها باید به‌طور هدفمند و در بخش‌هایی که قابلیت ایجاد تحرک را در سایر بخش‌های اقتصادی دارند صورت پذیرد. شناسایی بخش‌های کلیدی از ابتدا در چارچوب الگوهای داده- ستانده صورت می‌گیرد. در این زمینه روش‌های متعدد و متنوعی ازجمله روش چنری- واتانابه، راس‌موسن، روش فرضیه حذف، روش ریشه‌های مشخصه، روش تلفیقی داده- ستانده و اقتصادسنجی، منطق فازی و غیره استفاده شده که عمدتاً بر محاسبه پیوندهای پسین و پیشین بخش‌های اقتصادی مبتنی است. در این میان، منطق فازی علاوه بر استفاده از پیوندهای بین بخشی، قابلیت استفاده از سایر شاخص‌های اقتصادی بخش‌ها ازجمله تولید، صادرات، اشتغال و غیره را برای تعیین بخش‌های کلیدی دارد. همچنین، در منطق فازی می‌توان از داده‌های کیفی ازجمله سطح فناوری بخش‌ها در شناسایی بخش‌های کلیدی استفاده کرد. این روش، بخش‌های اقتصادی را که بیشتر با هم شباهت دارند در یک خوشه قرار می‌دهد و با توجه به ویژگی‌های هر خوشه می‌توان خوشه‌ای را که بخش‌های کلیدی در آن قرار دارد شناسایی کرد. دیاز و همکارانش[1] (۲۰۰۶) برای اولین بار در سال ۲۰۰۶ روشی را ارائه کردند که تلفیقی از تحلیل داده- ستانده و خوشه‌بندی صنعتی بود، در این روش در کنار استفاده از روابط بین‌بخشی، از سایر ویژگی‌های اقتصادی بخش‌ها مثل تولید، صادرات، اشتغال، دستمزد پرداختی و سطح فناوری بخش‌ها برای شناسایی بخش‌های کلیدی استفاده کردند. روش استفاده‌شده در این پژوهش نیز تلفیق شاخص‌های بین بخشی با سایر شاخص‌های اقتصادی بخش‌ها، ازجمله تولید، صادرات، اشتغال و میانگین رشد سالانه تولید، صادرات و اشتغال و استفاده از منطق فازی به‌منظور خوشه‌بندی و شناسایی بخش‌های کلیدی اقتصاد ایران است. در این روش می‌توان دید بهتری به بخش‌های مختلف اقتصاد بر مبنای شاخص‌های تعریف شده

1- Diaz et al. (2006)

داشت. خوشه‌بندی فازی حالت کلی‌تر خوشه‌بندی سخت است که بخش‌ها را بر اساس ویژگی‌های مشترک خوشه‌بندی می‌کند. در این روش هر بخش به بیش از یک خوشه با درجات عضویت مختلف تعلق دارد. این ویژگی خوشه‌بندی فازی، توان تحلیل خوشه‌ها را بر اساس اینکه کدام بخش‌ها در مرکز خوشه و کدام در مرزهای خوشه قرار دارند، بالاتر می‌برد.

در این مقاله قصد داریم با استفاده از خوشه‌بندی فازی و الگوریتم فازی ارائه‌شده هاثاوی و بزدک[1] با عنوان (Non-Euclidean RFCM) و تعریف متغیرهایی ازجمله میزان تولید، میزان صادرات، اشتغال هر بخش و غیره بخش‌های اقتصادی را خوشه‌بندی و بخش‌های کلیدی اقتصاد ایران را مشخص کنیم. البته پیش از استفاده از این الگوریتم، همان‌طور که دیاز و همکاران در مقاله خود درباره ضرورت شناسایی و جداسازی داده‌های پرت قبل از خوشه‌بندی می‌گوید با استفاده از روش‌های آماری شناسایی داده‌های پرت چندمتغیره، آن‌ها را شناسایی و به‌طور جداگانه خوشه‌بندی می‌کنیم سپس با توجه به ویژگی‌های هر خوشه، بخش‌های کلیدی اقتصاد شناسایی، و مقاله به‌صورت زیر سازمان‌دهی می‌شود. در ابتدا چارچوب نظری و مدل تحقیق ارائه، سپس پیشینه تحقیق آورده بیان می‌گردد. به دنبال آن داده‌های مقاله و محاسبات تجربی آن ذکر می‌شود و درنهایت مطالب خلاصه و جمع‌بندی می‌گردد.

۲- چارچوب نظری

نظریات رشد و توسعه پایه نظری اهمیت بخش‌های کلیدی اقتصاد هستند. دراین‌بین، سه نظریه شناخته‌شده رشد متوازن روزشتاین- رودن،[2] رشد نامتوازن هیرشمن[3] و قطب رشد[4] پرو و میردال[5] برای نشان دادن اهیمت بخش‌های کلیدی بررسی می‌شود.

روزشتاین- رودن نظریه رشد متوازن را که به نظریه فشار همه‌جانبه معروف است مطرح کرد. این نظریه بیان می‌کند یک بخش به‌تنهایی قادر به فراهم کردن توسعه اقتصادی نیست، بلکه اگر چندین بخش با بازدهی فزاینده و مرتبط به هم دست به تولید بزنند به‌گونه‌ای که هر یک تقاضایی برای محصول دیگری فراهم آورد، توسعه اقتصادی میسر خواهد شد (جهانگرد ۱۳۹۳: ۲۶۱).

1- Hathaway & Bezdek (1994)
2- Rosenstein-Rodan
3- Hirschman
4- Growth Poles
5- Myrdal

روزشتاین- رودن در این خصوص بیان می‌کند که قرار دادن کشوری در راه توسعه مداوم اقتصادی، شبیه قرار دادن یک هواپیما در باند پرواز است. هواپیما قبل از پرواز باید به‌قدری سرعت بگیرد که آماده پرواز شود. این امر شرط لازم برای موفقیت در پرواز است. به عقیده وی، با توسعه تدریجی و اعمال سیاست‌های اقتصادی به‌صورت گام‌به‌گام نمی‌توان اقتصاد را به‌طور موفقیت‌آمیز در خط رشد اقتصادی قرار داد. برای رشد مداوم اقتصادی به حداقل سرمایه‌گذاری نیاز است که باید به‌طور همه‌جانبه و یکباره صورت گیرد. در غیر این صورت هواپیما بر روی باند خواهد ماند و کماکان عرض باند را طی می‌کند و بار دیگر به‌جای خود باز می‌گردد (قره‌باغیان ۱۳۷۲: ۲۸۲-۲۸۳)

اگرچه این نظریه برای کشورهای توسعه‌یافته اجراشدنی بود، به دلیل محدودیت منابع سرمایه‌ای و انسانی و زیرساختی، این نظریه برای کشورهای درحال‌توسعه قابلیت پیاده‌سازی و اجرا نداشت؛ بنابراین هیرشمن، منتقد نظریه رشد متوازن، نظریه رشد نامتوازن را مطرح کرد.

طرفداران نظریه رشد نامتوازن عنوان می‌کنند این نظریه و نظریه همه‌جانبه به سرمایه‌گذاری‌های وسیع و هم‌زمان نیازمند است، درحالی‌که مشکل اصلی کشورهای توسعه‌نیافته کمبود سرمایه است. از سوی دیگر، با اجرای هم‌زمان سرمایه‌گذاری‌ها و طرح‌های مختلف، مشکل برنامه‌ریزی به وجود می‌آید و ممکن است در اثر اشتباه در برنامه‌ریزی و تخصیص نادرست منابع، از کارایی آن‌ها کاسته شده و بسیاری از منابع تلف شود؛ از این‌رو باید سرمایه‌های موجود و در دسترس را برای بخش‌ها یا صنایعی در نظر گرفت که بتواند نقش محرک را برای بخش‌ها یا صنایع دیگر ایفا کند؛ یعنی منابع لازم برای سرمایه‌گذاری در بخش‌های دیگر با منافع حاصل از سرمایه‌گذاری در بخش‌های پیشرو فراهم می‌شود و از این طریق صرفه‌جویی‌ها و توسعه اقتصادی به دست می‌آید (همان: ۲۶۳-۲۶۴)

نکته‌ای که در خصوص نظریه هیرشمن وجود دارد، استوار بودن آن بر تصمیم‌گیری طرح‌های اولویت‌دار است. درواقع، تصمیم‌گیری که موجب بیشترین اثر سرمایه‌گذاری (مستقیم القایی) در بخش‌های دیگر می‌شود، باید مدنظر قرار گیرد. همچنین درک میزان اهمیت وابستگی و پیوند متقابل میان فعالیت‌ها راهنمای خوبی برای سنجش بیشترین و مؤثرترین تصمیم‌گیری القایی است. به عقیده هیرشمن، این موضوع را باید در قالب پیوندها یعنی پیوندهای پسین و پیشین جست‌وجو کرد.

درنهایت تفاوت کارکرد دو نظریه رشد متوازن و رشد نامتوازن در میزان توسعه‌یافتگی و منابع موجود کشورهاست. در این زمینه پل استریتن عنوان می‌کند که باید بین رشد نامتوازن به‌منزله یک

روش، و رشد متوازن به‌عنوان یک هدف تمایز قائل شد. درواقع، زیربنای اساسی این دو نظریه منابع سرمایه و قدرت برنامه‌ریزی اقتصادی است. میزان دسترسی به هر کشور به این پارامترها، تعیین‌کننده جایگاه هر کشور در مسیر توسعه است؛ به‌گونه‌ای که اگر کشوری این پارامترها را نداشته باشد، بهتر است سیاست رشد نامتوازن را در پیش گیرد و در بلندمدت با فراهم شدن امکانات و شرایط لازم وارد مسیر رشد متوازن شود؛ زیرا رشد نامتوازن بر لزوم صرفه‌جویی در کاربرد منابع استوار است (گتاگ و سابرتا ۱۳۶۹: ۱۱۲).

نظریه قطب رشد که اقتصاد دانانی چون پرو و میردال مطرح کردند، بر دو اثر استوار است؛ یکی پیامدهای تمرکز و دیگری اثرات پخش، بدین‌صورت که رشد هم‌زمان در همه‌جا اتفاق نمی‌افتد، بلکه در نقاط یا قطب‌های توسعه رخ می‌دهد که قدرت جاذبه بالایی دارند (اثر تمرکز)، این نقاط، توسعه را در کانال‌هایی پخش می‌کنند که کل اقتصاد را تحت تأثیر قرار می‌دهد (اثر پخش). درواقع این مفهوم ابداعی پرو برگرفته از ایده شومپیتر است که رشد را محصول مستقیم و غیرمستقیم نوآوری می‌داند؛ بنابرین نوآوری مؤسسات بزرگ عامل اصلی و اولیه پیشرفت اقتصادی است. نظریه شومپیتر و نظریه ارتباطات درونی و بین صنایع، دو سنگ‌بنای اصلی نظریه پرو است. بر اساس نظریه ارتباطات درونی، نوآوری‌های به وجود آمده در یک فعالیت با مؤسسات دیگر به‌طور فزاینده گسترش می‌یابد. صنایع جدید نیز عمدتاً بخش‌هایی را شامل می‌شود که نوآوری زیادی دارند و در مقایسه با بخش‌های دیگر با رشد بیشتری همراه هستند، همچنین ارتباطات تنگاتنگی با بخش‌های پیشین و پسین خود برقرار می‌کنند (کلانتری ۱۳۸۰: ۷۱).

خلاصه نظریه قطب رشد را می‌توان در ترکیبی از فروض، اصول و راهبردهای زیر ترسیم کرد:

صرفه‌جویی‌های داخلی مانند کاهش هزینه واحد تولید با افزایش مقیاس تولید و نیز کاهش هزینه واحد تولید با دسترسی به دانش فنی تولید مربوط به بزرگ شدن واحدهای تولیدی می‌انجامد.

صرفه‌جویی‌های خارجی شامل کاهش هزینه واحد تولید ناشی از دسترسی به خدمات یا ناشی از تقسیم هزینه‌های مشترک تولید در یک‌رشته فعالیت (تجمیع واحدهای تولیدی) را تشویق می‌کند.

تلفیق دو مبحث بالا، به راهبردهای فضایی گوناگونی در توسعه منطقه‌ای منتهی می‌شود؛ به‌بیان‌دیگر، نظریه توسعه منطقه‌ای در این رویکرد با این استدلال توجیه‌پذیر می‌شود که روند توسعه برحسب ماهیت، غیرمتوازن آغاز شده و کارایی اقتصادی با تمرکز خدمات زیربنایی و فعالیت‌های مولد در کنار یکدیگر به دست می‌آید (جهانگرد : ۲۶۸-۲۶۹).

علیرغم وجود اتفاق‌نظری اساسی در مورد اهمیت پیوندهای پسین و پیشین در بین بخش‌های اقتصادی به‌منظور گسترش تحرک رشد اقتصادی بخش‌ها با استفاده از الگوی داده- ستانده، توافق کلی در مورد راه‌های تشخیص بخش‌های کلیدی در متون اقتصادی وجود ندارد. روش‌های متعددی ازجمله روش چنری- واتانابه،[1] راس‌موسن،[2] روش فرضیه حذف،[3] روش ریشه‌های مشخصه،[4] روش کشش‌های داده- ستانده، روش پیوندهای خالص و ناخالص،[5] روش تلفیقی داده- ستانده و اقتصادسنجی- تحلیل پوششی داده‌ها، روش ترکیبی داده- ستانده و منطق فازی[6] و روش تصادفی داده- ستانده در متون اقتصادی مطرح شده است؛ ولی اگرچه روش‌های زیادی برای شناسایی بخش‌های کلیدی وجود دارد تمامی این روش‌ها با تعریف پیوندهای پسین و پیشین مرتبط هستند که این مقاله از روش خوشه‌بندی فازی و داده- ستانده استفاده می‌کند.

۲-۱- تلفیق داده- ستانده و منطق فازی

این روش که دیاز و همکارانش (۲۰۰۶) برای شناسایی بخش‌های کلیدی اقتصاد استفاده کرده‌اند، ترکیبی از روش‌های داده- ستانده و استفاده از خوشه‌بندی صنعتی[7] است. در این روش علاوه بر استفاده از شاخص‌های بین بخشی ازجمله ضریب گش،[8] شاخص پراکندگی ضریب[9] گش، درجه انسجام[10] و شاخص روابط متقابل[11] بین بخش‌ها از سایر شاخص‌ها ازجمله ظرفیت نوآوری، پتانسیل و پویایی صادرات و سایر شاخصه‌های بخش‌ها برای ارتقا شناسایی بخش‌های کلیدی استفاده کردند (دیاز و همکاران[12] ۲۰۰۶: ۳۰۰).

بخش‌ها با استفاده از الگوریتم‌های خوشه‌بندی در تعداد مشخصی از خوشه، خوشه‌بندی می‌شوند که هر عضو بیشترین شباهت را با اعضای خوشه خود و بیشترین

1- Chenery & Watanabe
2- Rasmusen
3- Extraction Hypothesis
4- Eigen Value
5- Net Linkages
6- Fuzzy Logic
7- Industrial Clustering
8- Gosh's Multiplier
9- Multipliers' Variation Coefficient
10- Cohesion Grade
11- Interrelationship Index
12- Diaz et al.

تفاوت را با اعضای سایر خوشه‌ها دارد. در روش‌های خوشه‌بندی آماری، هر عضو تنها به یک خوشه مشخص تعلق دارد، اما در خوشه‌بندی فازی، به‌دلیل نبود مرزهای تیز[1] هر بخش ممکن است به‌طور پیوسته به چند خوشه تعلق داشته باشد و درجه عضویت بین صفر و یک به جای اختصاص هر بخش به هر خوشه استفاده می‌شود. با این روش هر بخش، با درجه‌ای متفاوت به هر خوشه تعلق دارد. بعد از خوشه‌بندی بخش‌های اقتصاد، با توجه به ویژگی‌های هر خوشه می‌توان بخش‌های کلیدی اقتصاد را شناسایی کرد (همان: ۳۰۱).

یکی از ویژگی‌های این روش در مقایسه با سایر روش‌ها این است که می‌توان علاوه بر متغیرهای کمی، از متغیرهای اسمی[2] هم در شناسایی بخش‌های کلیدی استفاده کرد. با این ویژگی می‌توان متغیرها و شاخص‌های بیشتری را برای شناسایی بخش‌های کلیدی به‌کار برد. از دیگر ویژگی‌های این روش قابلیت استفاده از شاخص‌های متعدد در تعیین بخش‌های کلیدی است (همان: ۳۰۱).

الف) روش‌های خوشه‌بندی

روش‌های خوشه‌بندی به دو دسته کلی سخت و فازی، و خوشه‌بندی سخت نیز به دو دسته سلسله‌مراتبی[3] و تفکیکی[4] تقسیم می‌شوند. روش‌های سلسله‌مراتبی هم به دو دسته ادغامی[5] و شکافتی[6] دسته‌بندی می‌شوند. روش خوشه‌بندی تفکیکی برای حل مسائل بزرگ (مسائلی که یا تعداد اشیاء یا تعداد شاخص‌ها زیاد است) کاربرد دارد. روش‌های خوشه‌بندی فازی نیز شکل گسترش‌یافته روش‌های سخت است که از مجموعه‌های فازی در خوشه‌بندی استفاده می‌کنند.

ب) خوشه‌بندی فازی

روش‌های خوشه‌بندی فازی، شکل گسترش‌یافته روش‌های خوشه‌بندی سخت (قطعی)

1- Sharp Boundaries
2- Nominal Variable
3- Hierarchical
4- Partitional
5- Agglomerative
6- Divisive

است. روش‌های فازی بسیار متنوع هستند و هرکدام از آن‌ها تابع عضویت و تابع هدفی را به کار می‌برند. برخی از این روش‌ها عبارت‌اند از: روش‌های فازی k میانگین، روش‌های فازی k مد و روش‌های فازی c میانگین.

روش فازی k میانگین،[1] شکل گسترش‌یافته روش k میانگین است. برای مجموعه اشیایی با n شیء و k خوشه، اگر X_j را بردار داده‌ها تعریف کنیم، الگوریتم فازی k میانگین به دنبال حداقل کردن تابع هدف زیر است

$$J_q(U,V)=\sum_{j=1}^{n}\sum_{i=1}^{k}u_{ij}^{q}d^2(X_j,V_i) \qquad (۱)$$

که در آن، u_{ij} درجه عضویت شیء j در خوشه i ام، d فاصله، V_i مرکز خوشه i و q عدد صحیح بزرگ‌تر از یک است که میزان فازی بودن را در خوشه‌بندی کنترل می‌کند (هرچقدر این پارامتر کمتر باشد میزان فازی بودن نیز کاهش می‌یابد و به‌طورمعمول مقدار این پارامتر برابر ۲ است). این تابع هدف به‌طور مستقیم قابلیت حداقل‌سازی ندارد؛ بنابراین از الگوریتم‌های تکرار برای حداقل سازی این تابع هدف استفاده می‌شود (مؤمنی ۱۳۹۰: ۲۲۴-۲۲۵).

روش فازی k مد[2] نیز مشابه روش فازی k میانگین است با این تفاوت که به جای میانگین، از مد داده‌ها با فراوانی داده‌ها استفاده می‌شود. این روش، روش k میانگین را که صرفاً برای داده‌های عددی کاربرد داشت گسترش داد و برای مجموعه داده‌های بیشتری در جهان کاربردی کرد (هوانگ و میشل[3] ۱۹۹۹: ۴۴۶).

روش خوشه‌بندی فازی c میانگین،[4] مشابه روش‌های دیگر خوشه‌بندی فازی، تابع هدفی است که آن را با استفاده از الگوریتم تکرار حداقل می‌کند. تابع هدف الگوریتم به‌صورت معادله (2) است

$$J_m(U,V;X)=\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m\left\|x_k-v_i\right\|^2 \qquad (۲)$$

1- Fuzzy k-means
2- Fuzzy k-modes
3- Huang and Michael
4- Fuzzy c-means

که در آن v_i مرکز خوشه i ام است که با معادله زیر در فرایند تکرار محاسبه می‌شود.

$$v_i^{(t)} = \frac{\sum_{k=1}^{n} (u_{ik}^{(t)})^m x_k}{\sum_{k=1}^{n} (u_{ik}^{(t)})^m} \qquad (۳)$$

در این معادله، m مقدار ثابت فازی بودن، t شمارنده حلقه، c تعداد خوشه‌ها که از پیش تعیین‌شده است و u_{ik} درجه عضویت عنصر i ام در خوشه k ام است که بعد از محاسبه مقدار $v_i^{(t)}$، توسط معادله (۳) محاسبه و فرایند تکرار آن‌قدر تکرار می‌شود تا شرط $\left|U^{(t+1)} - U^{(t)}\right| < \varepsilon$ برقرار گردد یا تعداد تکرارهای از پیش تعیین‌شده تمام شود (بزدک و همکاران[1] ۱۹۸۴: ۱۹۳).

$$u_{ik}^{(t+1)} = \left[\sum_{j=1}^{c} \left(\frac{\left\| x_k - v_i^{(t)} \right\|^2}{\left\| x_k - v_j^{(t)} \right\|^2} \right)^{\frac{2}{m-1}} \right]^{-1} \qquad (۴)$$

الگوریتم‌های متفاوتی از روش فازی c میانگین تولید شده است که می‌توان به FCM[2]، RFCM[3] و NERFCM اشاره کرد. در الگوریتم OFCM (خوشه‌بندی فازی c میانگین با استفاده از ماتریس اشیاء)، فرایند خوشه‌بندی با استفاده از ماتریس اشیاء اجرا می‌شود که ازاین‌جهت در خوشه‌بندی داده‌های عدم تشابه ضعف دارد. روش RFCM (خوشه‌بندی فازی c میانگین با استفاده از ماتریس عدم تشابه یا مشابهت)، فرایند خوشه‌بندی را با استفاده از ماتریس مشابهت یا عدم تشابه انجام می‌دهد که در مقایسه با روش OFCM قابلیت بیشتری دارد؛ زیرا ماتریس اشیاء را می‌توان به‌راحتی به ماتریس عدم تشابه (یا مشابهت) تبدیل کرد. باوجود این، برای اجرای این الگوریتم باید شرط اقلیدسی بودن ماتریس عدم مشابهت[4] وجود داشته باشد، درحالی‌که بسیاری از داده‌های

1- Bezdek et al.
2- Object Fuzzy c-means
3- Relational Fuzzy c-means
۴- شرط اقلیدسی بودن به این معنی است که n نقطه در فضای R^{n-1} وجود داشته باشد که فاصله اقلیدسی این نقاط دقیقا برابر با ماتریس عدم تشابه داده‌شده، برابر باشد.

مـوردنظر ایـن شـرط را ندارنـد. الگـوریتم NERFCM ایـن مشـکل را درون فراینـد خوشه‌بندی، مشابه به استفاده از یک تبدیل گسترش‌دهنده[1] برای ماتریس عدم تشابه، حل می‌کند؛ به‌گونه‌ای که می‌توان این الگوریتم را برای محـدوده وسـیعی از داده‌هـا اسـتفاده کرد (هاثاوی و بزدک[2] ۱۹۹۴: ۴۳۰-۴۲۹).

روش استفاده‌شده در این پژوهش، خوشه‌بندی فازی بخش‌ها با استفاده از شاخص‌هـایی است که برای هر بخش تعریف می‌شود. در ادامه ابتدا به مراحل آمـاده‌سـازی داده‌هـا بـرای استفاده در الگوریتم فازی پرداخته می‌شود. این مراحل شامل آمـاده‌سـازی و نرمـال کـردن ماتریس داده‌ها، شناسایی و جداسازی داده‌های پرت و محاسبه مـاتریس عـدم تشـابه اسـت. پس از آن الگوریتم خوشه‌بندی فازی RFCM و NERFCM که هاثاوی و بـزدک (۱۹۹۰) ارائـه کـرده‌انـد، توضـیح داده مـی‌شـود و درنهایـت تسـت سـیلووت بـرای اعتبارسـنجی خوشه‌بندی و تعیین تعداد صحیح خوشه‌ها آورده می‌شود.

ج) آماده‌سازی داده‌ها برای استفاده در الگوریتم خوشه‌بندی

الگوریتم خوشه‌بندی موردنظر، از ماتریس عدم تشابه برای خوشه‌بندی استفاده می‌کنـد؛ بنابراین پیش از اجرای الگوریتم به آماده‌سازی ماتریس عدم تشابه نیاز اسـت. اگـر بـرای n شیء p شاخص که ویژگی‌های اشـیاء را توضـیح مـی‌دهنـد تعریـف کنـیم، مـاتریس اشـیاء به‌صورت زیر تعریف می‌شود.

$$O = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

با توجه به اینکه شاخص‌ها مقیاس‌های متفاوتی دارند، با استفاده از معادلـه زیـر مـاتریس داده‌ها را نرمال می‌کنیم (هر سطر ماتریس داده‌ها با فرمول زیر نرمال می‌شود):

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{ad_j} \qquad (۵)$$

1- Spreading Transformation
2- Hathaway and Bezdek

که در آن، $\bar{x}_i$ و ad_i با معادلات $\bar{x}_i = \frac{\sum_{j=1}^{p} x_{ij}}{p}$ و $ad_i = \frac{\sum_{j=1}^{p} \left| x_{ij} - \bar{x}_i \right|}{p-1}$ محاسبه[۱] می‌شوند. به این دلیل از انحراف مطلق به‌جای انحراف معیار استفاده می‌شود که داده‌های پرت تأثیر کمتری بر فرایند نرمال کردن می‌گذارد و در فرایند شناسایی داده‌های پرت، داده‌های پرت شناسایی می‌شوند.

د) شناسایی داده‌های پرت چندمتغیره

روش‌های متعدد شناسایی داده‌های پرت تک‌متغیره و چندمتغیره برای شناسایی داده‌های پرت، اعم از پارامتری و ناپارامتری وجود دارد. در این پژوهش از روش شناسایی داده‌های پرت چندمتغیره PCS[۲] که وکیلی و اشمیت[۳] (۲۰۱۴) ارائه کرده‌اند استفاده می‌کنیم. این روش در مقایسه با سایر روش‌های مشابه از جمله MCD،[۴] MVE[۵] و SDE[۶] تحت تأثیر حضور داده‌های پرت قرار نمی‌گیرد. اگرچه این روش محدودیتی در نسبت تعداد اشیاء به تعداد شاخص‌ها دارد، تعداد اشیاء (n) باید حداقل بیشتر از پنج برابر تعداد شاخص‌های تعریف‌شده (p) باشد تا بتوان از این الگوریتم برای شناسایی داده‌های پرت استفاده کرد.

روش PCS که معیار دورافتادگی‌ای[۷] را تعریف می‌کند زمانی حداقل می‌شود که زیرمجموعه H_m بیشترین هم‌پوشانی را با زیرمجموعه H_{mk} (زیرمجموعه بهینه‌ای است که اشیای آن بیشترین تجانس را دارد) داشته باشد.

$$I(H_m, a_{mk}) := \log \frac{\underset{i \in H_m}{ave}\, d^2_{p,i}(a_{mk})}{\underset{i \in H_{mk}}{ave}\, d^2_{p,i}(a_{mk})} \qquad (۶)$$

که در آن $d^2_{p,i}(a_{mk})$ فاصله متعامد[۸] x_i از a_{mk} است که توسط معادله زیر محاسبه می‌شود.

۱- در تابع نرمال استاندارد بجای s_i از ad_i استفاده می‌شود (زکی و میرا ۲۰۱۴: ۵۳-۵۲)

2- Projection Congruent Subset
3- Vakili & Schmitt
4- Minimum Covariance Determinant
5- Minimum Volume Ellipsoid
6- Stahel–Donoho Estimator
7- Outlyingness Index
8- Orthogonal Distance

$$d^{2}_{p,i}(a_{mk}) = \frac{(x_i^{'} a_{mk} - 1)^2}{\|a_{mk}\|^2} \qquad (۷)$$

و بردار a_{mk} به صورتی تعریف می‌شود که اگر A_{mk} ماتریس از p مشاهده باشد (این مشاهدات با الگوریتم انتخاب می‌شود)، آن‌گاه $A_{mk} a_{mk} = 1_p$ شود (وکیلی و اشمیت، ۲۰۱۴، ص ۵۶-۵۷)[۱]

هـ) محاسبه ماتریس عدم تشابه

با استفاده از فاصله اقلیدسی می‌توان ماتریس اشیاء را به ماتریس فاصله یا عدم تشابه تبدیل کرد. اگر x_i و x_j دو نقطه در فضای $\mathbb{R}^P$ باشند، آنگاه فاصله اقلیدوسی بین این دونقطه به‌صورت زیر محاسبه می‌شود.

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \qquad (۸)$$

با محاسبه این فاصله برای تمامی اشیاء ماتریس داده، این ماتریس ($n \times p$) به ماتریس عدم تشابه ($n \times n$) تبدیل می‌شود.

$$D = \begin{pmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{pmatrix}_{n \times n}$$

که در آن سه شرط زیر برقرار است:

$$\begin{cases} d_{ii} = 0 \\ d_{ij} \geq 0 \quad i \neq j \\ d_{ij} = d_{ji} \quad i \neq j \end{cases} \qquad (۹)$$

(هاثاوی و بزدک ۱۹۹۴، ص ۴۲۹)

و) الگوریتم خوشه‌بندی فازی RFCM و NERFCM

همان‌طور که گفته شد روش خوشه‌بندی در این پژوهش، الگوریتم خوشه‌بندی فازی

۱- شاخص دورافتادگی و داده‌های پرت با الگوریتم FastPCS که وکیلی و اشمیت (۲۰۱۴) ارائه کرده‌اند محاسبه می‌شود.

NERFCM است که هاثاوی و بزدک[1] (۱۹۹۴) ارائه کرده‌اند، این الگوریتم که اشیاء را با استفاده از ماتریس فاصله یا عدم تشابه خوشه‌بندی می‌کند، شکل گسترش‌یافته الگوریتم RFCM است که مشکل اقلیدسی بودن ماتریس عدم تشابه را حل می‌کند. پیش از ارائه این الگوریتم، مفاهیم موردنیاز همچنین الگوریتم RFCM توضیح داده می‌شود. مجموعه M_{fcn}، یک مجموعه فازی از تمامی ماتریس‌های $U=[U_{ik}]\in\mathbb{R}^{c\times n}$ است که سه شرط زیر را ارضا می‌کند:

$$\begin{cases} U_{ik}\in[0,1] & for\ 1\le i\le c\ and\ 1\le k\le n \\ U_{1k}+\cdots+U_{nk}=1 & for\ 1\le k\le n \\ U_{i1}+\cdots+U_{in}>0 & for\ 1\le i\le n \end{cases} \qquad (۱۰)$$

ماتریس $U_{c\times n}$ درجه عضویت n شیء در c خوشه را مشخص می‌کند. به‌طور مثال ماتریس فازی سه بخشی زیر نشان می‌دهد که شیء o_1 با درجه عضویت بالایی به خوشه c_1 تعلق دارد ($U_{11}=0.95$). اشیاء o_2 و o_3 به خوشه دوم تعلق دارند و شیء o_4 نسبتا به دو خوشه دو و سه تعلق دارد ($U_{24}=0.47, U_{34}=0.51$).

$$U=\begin{pmatrix} 0.95 & 0.11 & 0.05 & 0.02 \\ 0.03 & 0.78 & 0.81 & 0.47 \\ 0.02 & 0.11 & 0.14 & 0.51 \end{pmatrix}$$

مسئله خوشه‌بندی فازی، پیدا کردن ماتریس بهینه $U\in M_{fcn}$ با حداقل کردن شرط زیر است:

$$J_m(U,V)=\sum_{k=1}^{n}\sum_{i=1}^{c}(U_{ik})^m\left\|x_{ik}-v_i\right\|^2 \qquad (۱۱)$$

که در آن c تعداد خوشه‌ها، $m>1$ پارامتر کنترل مقداری فازی بودن، $v_i\in\mathbb{R}^s$ میانگین خوشه c_i و $V\in\mathbb{R}^{c\times s}$ ماتریسی است که هر سطر آن v_i^T است. (همان، ص ۴۲۹)

ز) الگوریتم RFCM

الگوریتم خوشه‌بندی فازی RFCM با استفاده از تکرار، اشیایی را که با ماتریس عدم

1- Hathaway & Bezdek

تشابه $D=\left[D_{ij}\right]=\left[\left\|x_i-x_j\right\|^2\right]$ مشخص می‌شوند به‌صورت زیر خوشه‌بندی می کند ($\|.\|$ فاصله اقلیدسی است):

تعیین مقادیر c ($2\le c<n$) و m ($m>1$) و مقداردهی اولیه به ماتریس $U^{(0)}\in M_{fcn}$ و اجرای حلقه برای $r=1,2,...$

محاسبه بردارهای $v_i=v_i^{(r)}$ با استفاده از $U=U^{(r)}$ و معادله زیر برای $1\le i\le c$:

$$v_i=\frac{(U_{i1}^m,U_{i2}^m,\cdots,U_{in}^m)}{(U_{i1}^m+U_{i2}^m+\cdots+U_{in}^m)} \qquad (۱۲)$$

محاسبه d_{ik} برای $1\le i\le c$ و $1\le k\le n$ با استفاده از معادله (۱۳)

$$d_{ik}=(Dv_i)_k-\frac{(v_i^T Dv_i)}{2} \qquad (۱۳)$$

به روز رسانی ماتریس $U^{(r)}$ به $U=U^{(r+1)}\in M_{fcn}$ برای هر $k=1,\cdots,n$ به این صورت که:

اگر برای تمام i ها، $d_{ik}>0$ بود، آنگاه:

$$U_{ik}=\frac{1}{(\frac{d_{ik}}{d_{1k}}+\frac{d_{ik}}{d_{2k}}+\cdots+\frac{d_{ik}}{d_{ck}})^{\frac{1}{m-1}}} \qquad (۱۴)$$

در غیر این صورت، حداقل یکی از d_{ik} ها برابر صفر است، آنگاه:

مقدار U_{ik} را به گونه‌ای که در بازه [0,1] باشد و $(U_{1k}+\cdots+U_{ck})=1$ قرار دهید.

اگر $\left\|U^{(r+1)}-U^{(r)}\right\|\le\varepsilon$ برقرار بود، حلقه تکرار خاتمه می‌یابد، در غیر این صورت مقدار شمارنده r را یکی افزایش داده و به مرحله دوم برمی‌گردد.

این الگوریتم تا جایی که شرط حلقه برقرار شود، تکرار می‌گردد و درنهایت ماتریس بهینه‌شده U به‌منزله نتیجه خوشه‌بندی ارائه می‌شود. درصورتی‌که ماتریس عدم تشابه نااقلیدسی (دلخواه) باشد، این روش به خوشه‌بندی صحیح اشیاء قادر نیست. می‌توان

ماتریس عدم تشابه را با استفاده از تبدیل گسترش بتا[1] به شکل اقلیدسی درآورد، اما می‌توان به‌جای محاسبات زمان‌بر، از روش NERFCM استفاده کرد که ماتریس عدم تشابه را درون الگوریتم به شکل اقلیدسی درمی‌آورد (همان: ۴۳۱)

ح) الگوریتم NERFCM

این الگوریتم به الگوریتم RFCM شبیه است، با این تفاوت که هر ماتریس عدم تشابه دلخواهی که شرایط (۹) را مهیا کند، در این الگوریتم کاربرد دارد.

تعیین مقادیر c ($2 \le c < n$) و m ($m > 1$) و مقداردهی اولیه به ماتریس $U^{(0)} \in M_{fcn}$، $\beta = 0$ و اجرای حلقه برای $r = 1, 2, \ldots$

محاسبه بردارهای $v_i = v_i^{(r)}$ با استفاده از $U = U^{(r)}$ و معادله زیر برای $1 \le i \le c$:

$$v_i = \frac{(U_{i1}^m, U_{i2}^m, \cdots, U_{in}^m)}{(U_{i1}^m + U_{i2}^m + \cdots + U_{in}^m)} \qquad (۱۵)$$

محاسبه d_{ik} برای $1 \le i \le c$ و $1 \le k \le n$ با استفاده از معادله (۱۶)

$$d_{ik} = (Dv_i)_k - \frac{(v_i^T D v_i)}{2} \qquad (۱۶)$$

اگر d_{ik} برای هر i و k کمتر از صفر باشد آنگاه:

$$\Delta\beta = \max\left\{-2 * d_{ik} / (\|v_i - e_k\|^2)\right\} \qquad (۱۷)$$

$$d_{ik} \leftarrow d_{ik} + (\frac{\Delta\beta}{2}) * \|v_i - e_k\|^2 \quad for\ 1 \le i \le c\ \&\ 1 \le k \le n \qquad (۱۸)$$

$$\beta \leftarrow \beta + \Delta\beta \qquad (۱۹)$$

به‌روز رسانی ماتریس $U^{(r)}$ به $U = U^{(r+1)} \in M_{fcn}$ برای هر $k = 1, \cdots, n$ به این صورت که:

اگر برای تمام i ها، $d_{ik} > 0$ بود، آنگاه:

1- β-Spread Transformation

$$U_{ik} = \frac{1}{(\frac{d_{ik}}{d_{1k}} + \frac{d_{ik}}{d_{2k}} + \cdots + \frac{d_{ik}}{d_{ck}})^{\frac{1}{m-1}}} \qquad (۲۰)$$

در غیر این صورت، اگر $d_{ik} > 0$ باشد، آنگاه $U_{ik} = 0$ و $U_{ik} \in [0,1]$ و $(U_{1k} + \cdots + U_{ck}) = 1$

اگر $\left\| U^{(r+1)} - U^{(r)} \right\| \leq \varepsilon$ برقرار بود، حلقه تکرار خاتمه می‌یابد، در غیر این صورت مقدار شمارنده r را یکی افزایش داده و به مرحله دوم برمی‌گردد (همان: ۴۳۳).

ط)آزمون اعتبار خوشه‌بندی و تعیین تعداد خوشه‌ها (c)

اعتبار خوشه‌بندی و تعیین تعداد صحیح خوشه‌ها با استفاده از معیار سیلووت فازی[1] مشخص می‌شود که کامپلو و هراشکا[2] (۲۰۰۶) ارائه داده‌اند[3] این معیار که حالت گسترش داده‌شده معیار سیلووت سخت[4] است، به‌صورت زیر تعریف می‌شود:

اگر شیء $j \in \{1,2,\cdots,n\}$ در خوشه $p \in \{1,\cdots,c\}$ قرار بگیرد (در خوشه‌بندی سخت به این معنی است که شیء j به خوشه p در مقایسه با سایر خوشه‌ها نزدیک‌تر است و در خوشه‌بندی فازی به معنی این است که درجه عضویت شیء j در خوشه p نسبت به سایر خوشه‌ها بیشتر است) آن‌گاه اگر a_{pj} میانگین فاصله شیء j با سایر اشیاء درون خوشه p، همچنین d_{qj} میانگین فاصله شیء j با اشیاء درون خوشه q ($q \neq p$) و b_{qj} حداقل مقدار d_{qj} در بین خوشه‌های مختلف $q = 1,\cdots,c, q \neq p$ باشد (که فاصله شیء j با نزدیک‌ترین خوشه غیر از خوشه p را نشان می‌دهد)، سیلووت شیء j به‌صورت زیر تعریف می‌شود:

$$s_j = \frac{b_{pj} - a_{pj}}{\max\{a_{pj}, b_{pj}\}} \qquad (۲۱)$$

1- Fuzzy Silhouette (FS)

2- Campello and Heruschka

۳- معیارهای دیگری برای اعتبارسنجی و تعیین تعداد خوشه‌ها ازجمله AWCD، FHV، Xei-Beni index و ... وجود دارد اما معیار سیلووت از جهت سرعت محاسباتی نسبت به سایر روش‌ها ارجحیت دارد (کامپلو و هروشکا ۲۰۰۶).

4- Crisp Silhouette (CS)

هرچه مقدار s_j بیشتر باشد، به معنی تعلق بیشتر شیء j به خوشه p است. حال اگر $P=\left[u_{ij}\right]_{c\times n}$ ماتریس خوشه‌بندی فازی باشد، معیار سیلووت فازی به‌صورت معادله (۲۲) تعریف می‌شود.

$$FS=\frac{\sum_{j=1}^{n}(u_{pj}-u_{qj})^{\alpha}s_j}{\sum_{j=1}^{n}(u_{pj}-u_{qj})^{\alpha}} \qquad (۲۲)$$

که در آن، u_{pj} و u_{qj} به ترتیب اولین و دومین بزرگ‌ترین مقدار ستون j‌ام ماتریس خوشه‌بندی فازی و $\alpha\geq 0$ پارامتر کنترل‌کننده وزن است. هرچه مقدار معیار FS بیشتر باشد، اعتبار خوشه‌بندی بیشتر است (کامپلو و هراشکا ۲۰۰۶: ۲۸۶۲-۲۸۶۳).

با تکرار الگوریتم خوشه‌بندی فازی NERFCM برای تعداد خوشه $c=2,\cdots,n-1$ و حداکثر کردن معیار FS نسبت به مقدار c، می‌توان به تعداد خوشه صحیح رسید.

۳- پیشینه تحقیق

ازجمله مطالعات خارجی انجام‌شده در زمینه استفاده از خوشه‌بندی فازی در شناسایی بخش‌های کلیدی می‌توان به دو مقاله دیاز و همکاران (۲۰۰۶) و موریلاس و دیاز (۲۰۰۸) اشاره کرد. البته پیش از آن، زمانسکی[1] (۱۹۷۴) از فنون چندمتغیره آماری برای خوشه‌بندی صنعتی با استفاده از جدول داده- ستانده سال ۱۹۶۳ آمریکا استفاده کرد. ری و ماتیس[2] (۲۰۰۰) نیز با استفاده از روش ترکیبی و برای اولین بار با استفاده از یک روش خوشه‌بندی غیر سلسله مراتبی، به خوشه‌بندی صنعتی صنایع ایالت کالیفرنیا پرداختند، اما ترکیب روش‌های داده- ستانده و خوشه‌بندی فازی را دیاز ارائه کرده است.

دیاز و همکاران (۲۰۰۶) در مقاله خود تحت عنوان «یک راهکار خوشه‌بندی فازی برای بخش‌های کلیدی اقتصاد اسپانیا» به شناسایی بخش‌های اقتصادی اسپانیا با استفاده از

1- Czamanski
2- Rey & Mattheis

جدول داده- ستانده سال ۱۹۹۵ برای این کشور و با استفاده از خوشه‌بندی فازی پرداختند. آن‌ها در مقاله خود عنوان کردند رویکردهای متداول خوشه‌بندی این اجازه را به بخش‌ها نمی‌دهد که در چند گروه قرار بگیرند، درحالی‌که یک بخش می‌تواند از چند جنبه، بخش کلیدی البته با درجه‌ای متفاوت باشد؛ بنابراین بهتر است از منطق فازی برای خوشه‌بندی بخش‌های کلیدی استفاده شود؛ زیرا با می‌توان یک بخش را در چند گروه با درجه عضویت متفاوت قرار داد. درنهایت آن‌ها با استفاده از یک راهکار فازی چندمتغیره و با استفاده از معرفی سه گروه متغیر، به خوشه‌بندی بخش‌های اقتصادی اسپانیا اقدام کردند. آن‌ها بخش‌های اقتصادی اسپانیا را در سه گروه خوشه‌بندی کردند. گروه اول با عنوان گروه بخش‌های با پایه محلی[1] (با وزن اقتصادی[2] و یکپارچگی[3] بالا)، گروه دوم با یکپارچگی کمتر و متمایل به تقاضای داخلی و گروه سوم با گرایش تقاضای خارجی و با سطح تکنولوژی بالاتر. آن‌ها با در نظر گرفتن یک سری ملاحظات اضافی به این نتیجه رسیدند که گروه سوم بخش کلیدی اقتصاد اسپانیا، زمانی که به دنبال یک برنامه رشد اقتصادی هستیم باید در نظر گرفته شود.

موریلاس و دیاز[4] (۲۰۰۸) در مقاله‌ای با عنوان «بخش‌های کلیدی، خوشه‌بندی صنعتی، داده‌های پرت چندمتغیره» عنوان می‌کنند زمانی که بخش‌های اقتصادی را با استفاده از داده‌های چندبعدی خوشه‌بندی می‌کنیم، حضور داده‌های پرت چندمتغیره،[5] علاوه بر آنکه به کم شدن تعداد خوشه‌ها منجر می‌شود، نتایج خوشه‌بندی را نیز مخدوش می‌کند؛ بنابراین پیشنهاد می‌کنند ابتدا داده‌های پرت چندمتغیره شناسایی، سپس اقدام به خوشه‌بندی بخش‌ها شود. آن‌ها با استفاده از تخمین‌زن حداقل دترمینان ماتریس کواریانس[6] (MCD) و با معیار فاصله ماهالونوبیس[7] و تست Kruskal-Wallis، ۲۰ بخش از ۶۶ بخش را به‌عنوان داده

1- Local Base Sectors
2- Economic Weight
3- Integration
4- Morillas & Diaz
5- Multivariate Outliers
6- Minimum Covariance Determinant Estimator
7- Mahalanobis Distance

پرت شناسایی کردند. سپس با استفاده از الگوریتم خوشه‌بندی فازی NERFCM دو گروه داده‌های پرت و بخش‌های باقیمانده را به‌طور جداگانه خوشه‌بندی کردند، که گروه داده‌های پرت به ۹ خوشه و بخش‌های باقیمانده به ۶ خوشه، تقسیم‌بندی شدند. آن‌ها با استفاده از بررسی شاخص‌های تعریف‌شده هر خوشه، بخش‌های کلیدی اسپانیا را مشخص کردند.

در داخل بیشتر مطالعات تجربی انجام‌شده با استفاده از رویکرد داده- ستانده صرف و غیرتصادفی و غیرفازی بوده است. در مطالعات تشخیص بخش‌های کلیدی از رویکرد خوشه‌بندی فازی تاکنون استفاده نشده است، اما رویکرد داده- ستانده و فازی کاربرد دارد که مطالعه فنی ممتاز (۱۳۹۰) از این سبک است. وی در پژوهش خود با عنوان «شناسایی بخش‌های کلیدی اقتصاد ایران: رویکرد تلفیقی داده- ستانده و فازی» با استفاده از محاسبه فازی پیوندهای پسین و پیشین و تجزیه‌وتحلیل فازی آن‌ها به این نتیجه رسید که محاسبه و تحلیل فازی پیوندهای پسین و پیشین نتایجی مشابه با نتایج تحلیل معمول پیوندهای پسین و پیشین می‌دهد. او در پژوهش خود از جدول داده- ستانده سال ۱۳۸۰ که در ۲۰ بخش تجمیع شده استفاده کرد که درنهایت بخش‌های کلیدی اقتصاد ایران در هر دو روش تحلیل معمول پیوندهای پسین و پیشین و تحلیل فازی آن‌ها بخش‌های «آب، برق، گاز»، «محصولات ساخته‌شده از چوب»، «خمیر کاغذ و سایر محصولات کاغذی»، «محصولات شیمیایی همراه با لاستیک و پلاستیک»، «محصولات شیشه‌ای و کانی‌ها» و «فلزات و محصولات آن‌ها» بخش‌های کلیدی اقتصاد ایران هستند.

باید توجه داشت که مطالعات متعددی با رویکرد تلفیقی داده- ستانده و روش‌های تصادفی انجام شده است که نزدیک‌تر به رویکرد این مطالعه می‌باشد. دراین‌باره مطالعه جهانگرد و عاشوری (۱۳۸۹) یکی از آن‌هاست. این دو از تلفیق روش‌های تحلیل داده- ستانده و اقتصادسنجی، همچنین تحلیل پوششی داده‌ها برای شناسایی بخش‌های کلیدی اقتصاد ایران استفاده کردند. نتایج آن‌ها نشان‌دهنده خدمات محور بودن بخش‌های کلیدی اقتصاد ایران است، درحالی‌که روش سنتی داده- ستانده مؤید صنعت محور بودن اقتصاد از حیث بخش‌های کلیدی است. همچنین جهانگرد و کشت‌ورز (۱۳۹۰) در مقاله خود بر

اساس نظریه شبکه به مطالعه بخش‌های کلیدی اقتصاد ایران پرداختند. آن‌ها در مقاله خود با بهره‌گیری از مطالعه مونیز و همکاران (۲۰۰۸) سه معیار مرکزی جدید با عناوین اثرهای کلی، اثرهای میانی و اثرهای آنی درزمینه داده- ستانده، به‌منظور شناسایی بخش‌های کلیدی معرفی کردند. درنهایت با توجه به معیارهای معرفی‌شده، بیشترین اثرهای کلی طرف تقاضا به بخش‌های ارتباطات، تولید کاغذ و محصولات کاغذی، تولید چوب و محصولات چوبی، آب، برق و گاز مربوط است. مقاله جهانگرد و حسینی (۱۳۹۲) استفاده از رویکرد تحلیل تصادفی و روش برآورد فاصله‌ای و شبیه‌سازی مونت کارلو به تعیین بخش‌های اقتصادی به شکل تجربی معرفی شده است. آن‌ها در مقاله خود عنوان کردند به‌دلیل وجود خطاهای اندازه‌گیری و آماری بهتر است از روش تصادفی به جای روش غیر تصادفی استفاده شود. نتایج بررسی رویکرد غیر تصادفی راس‌موسن آن‌ها نشان می‌دهد که برای جلوگیری از بروز خطا در شناسایی بخش‌های کلیدی بهتر است از جدول تجمیع نشده به‌جای جدول تجمیع شده استفاده شود. هم‌چنین نتایج مقایسه دو روش غیر تصادفی و تصادفی راس‌موسن نیز نشان می‌دهد که تنها پنج بخش از شش بخش روش غیرتصادفی، در روش تصادفی بخش کلیدی هستند (به غیر از بخش کشاورزی، شکار، جنگلداری و ماهیگیری). همچنین چهار بخش که در روش تصادفی بخش کلیدی هستند درروش غیرتصادفی بخش کلیدی محسوب نمی‌شوند.

۴- داده‌ها

شاخص‌های استفاده‌شده مطابق چارچوب نظری و مدل تحقیق برای شناسایی بخش‌های کلیدی اقتصاد ایران، از سه بعد مختلف این بخش‌ها را بررسی می‌کند. گروه اول که شامل دو شاخص ضریب گش و معکوس پراکندگی ضریب گش می‌شود، روابط بین بخشی بخش‌های اقتصاد ایران را نشان می‌دهد. این دو شاخص که با استفاده از ماتریس معکوس گش (که از جدول داده- ستانده محاسبه می‌شود) محاسبه می‌شوند، برای بررسی میزان ارتباط بخش با سایر بخش‌ها انتخاب شده‌اند. هر چه مقدار این شاخص‌ها بیشتر باشد، نشان

از ارتباط بیشتر بخش با سایر بخش‌های دیگر دارد. ضریب گش، از جمع ستون‌های ماتریس معکوس گش برای هر بخش و پراکندگی ضریب گش، با استفاده از معادله (۲۳) از جدول داده- ستانده سال ۹۰ محاسبه شده‌اند.

$$CV_i^f = \frac{\sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(b_{ij} - \frac{1}{n}\sum_{j=1}^{n} b_{ij})^2}}{\frac{1}{n}\sum_{j=1}^{n} b_{ij}} \qquad (۲۳)$$

(دیاز و همکاران ۲۰۰۶: ۳۰۶).

که در آن b_{ij} عنصر ij ام ماتریس معکوس گش است. در این پژوهش از معکوس ضریب پراکندگی گش استفاده شده است.

گروه دیگر شاخص‌ها، شامل تولید، صادرات و اشتغال هر بخش است که از سه جنبه متفاوت، میزان وزن اقتصادی هر بخش را نشان می‌دهد؛ به عبارتی دیگر این شاخص‌ها میزان اهمیت هر بخش و نقش آن را در اقتصاد ایران بازگو می‌کند. دو شاخص تولید و صادرات هر بخش، با استفاده از جدول داده- ستانده سال ۱۳۹۰، به قیمت‌های جاری محاسبه شده‌اند. شاخص اشتغال هر بخش نیز با استفاده از سرشماری نفوس و مسکن سال ۱۳۹۰ به دست‌آمده است.

گروه آخر شاخص‌ها، روند تولید، صادرت و اشتغال هر بخش هستند. این گروه از شاخص‌ها به‌نوعی پتانسیل رشد هر بخش را در زمینه تولید، صادرات و اشتغال نشان می‌دهند. برای محاسبه این سه شاخص از دو جدول داده- ستانده ۱۳۸۵ و ۱۳۹۰ به قیمت ثابت استفاده شده است. شاخص رشد برای هر بخش، میانگین هندسی سالانه رشد تولید بخش از سال ۱۳۸۵ تا ۱۳۹۰ است. به دلیل اینکه مقادیر تولید در جداول داده- ستانده به قیمت جاری ارائه شده، با استفاده از شاخص قیمت ضمنی تولید مقادیر تولید در جداول به قیمت ثابت تبدیل گشته است. محاسبه رشد سالانه صادرات نیز به همین صورت محاسبه و استفاده شده است. رشد اشتغال با استفاده از دو سرشماری نفوس و مسکن سال‌های ۱۳۹۰ و ۱۳۸۵ و با گرفتن میانگین هندسی برای پنج سال محاسبه شده است.

آمارهای دیگر استفاده‌شده در این پژوهش، یک؛ جدول داده- ستانده بخش در بخش (۶۲ بخشی) سال ۱۳۸۵ است با فرض فنّاوری بخش و با واحد میلیارد ریال که مرکز پژوهش‌های مجلس شورای اسلامی ارائه کرد . دو؛ جدول داده- ستانده بخش در بخش (۷۱ بخشی) سال ۱۳۹۰ که مرکز پژوهش‌های مجلس شورای اسلامی، و با فرض فنّاوری بخش و با واحد میلیارد ریال، بوده است. به‌منظور هماهنگ‌سازی بخش‌های موجود در جداول ۱۳۸۵ و ۱۳۹۰، دو جدول در ۴۷ بخش تجمیع شده است. با توجه به جدول تجمیع شده ۴۷ بخشی و ۸ شاخص تعریف‌شده در بخش پیشین، ماتریس داده‌ها به‌صورت یک ماتریس ۸×۴۷ تشکیل می‌شود. برای یکسان‌سازی شاخص‌ها، هر ستون این ماتریس با استفاده از معادله (۵) نرمال می‌شود سپس با استفاده از الگوریتم FastPCS، داده‌های پرت شناسایی می‌شوند. داده‌ها پس از نرمال شدن، با استفاده از الگوریتم FastPCS برای شناسایی داده‌های پرت موردبررسی قرار گرفتند. نتایج زیر با استفاده از الگوریتم مذکور به دست آمد.

جدول ۱- شاخص دورافتادگی محاسبه‌شده برای بخش‌ها

بخش	Out. Index	بخش	Out. Index
نفت خام و گاز طبیعی	۳۶۱,۳۵	پست و مخابرات	۳,۳۶
ساخت کک، فراورده های نفتی و ...	۷۰,۷۹	آب	۳,۳۳
ساخت چوب و محصولات چوبی	۳۹,۲۹	سایر خدمات	۳,۰۷
عمده‌فروشی، خردهفروشی، تعمیر وسایل نقلیه و کالاها	۳۰,۶۶	حمل و نقل هوایی	۳,۰۱
خمیرکاغذ،کاغذومحصولات‌کاغذ یاوراق چاپی و کالاهای مربوط	۲۳,۸۴	دامداری، مرغداری، پرورش کرم ابریشم و زنبورعسل و شکار	۲,۹۵
توزیع گاز طبیعی	۱۷,۵	ساخت ماشین‌آلات و دستگاه‌های برقی طبقه‌بندی‌شده در جای دیگر	۲,۹۵
زراعت و باغداری	۱۶,۵۳	ماهیگیری	۲,۸۹
خدمات مستغلات	۱۵,۷۴	بانک و واسطه‌گری مالی	۲,۸۴
ساخت فلزات اساسی	۱۳,۷۶	سایر ساختمان‌ها	۲,۸۳

ادامه جدول ۱

بخش	Out. Index	بخش	Out. Index
حمل‌ونقل جاده‌ای	۱۲,۸۳	برق	۲,۷۳
ساخت رادیو و تلویزیون، دستگاه‌ها و وسایل ارتباطی	۱۰,۷۱	ساخت ابزار پزشکی، ابزار اپتیکی، ابزار دقیق و انواع ساعت	۲,۵۸
ساخت وسایل نقلیه موتوری، تریلر و ...	۱۰,۶۴	ساخت پوشاک، عمل‌آوری و رنگ کردن خز	۲,۷۴
حمل‌ونقل آبی	۱۰,۵۳	خدمات آموزش عمومی – فنی حرفه ای	۲,۴
سایر محصولات‌غذایی و آشامیدنی و محصولات از توتون و تنباکو	۱۰,۱۶	بهداشت و درمان	۲,۲۵
سایر معادن	۷,۲۹	حمل‌ونقل لوله‌ای	۲,۲
ساخت منسوجات	۶,۸۱	ساختمان‌های مسکونی	۲,۰۴
ساخت سایر محصولات کانی غیر فلزی	۴,۴۴	خدمات پشتیبانی و انبارداری	۱,۸۹
ساخت ماشین‌آلات و تجهیزات طبقه‌بندی نشده در جای دیگر	۳,۷۹	دباغی و پرداخت چرم و سایر محصولات چرمی	۱,۸۸
امور عمومی	۳,۷۵	آموزش عالی	۱,۸۷
هتل و خوابگاه و رستوران	۳,۵۷	ساخت محصولات فلزی فابریکی به جز ماشین‌آلات و تجهیزات	۱,۷۲
ساخت ماشین‌آلات دفتری، حسابداری و محاسباتی	۳,۴۷	تفریحی، فرهنگی، و ورزشی	۱,۴۷
بیمه	۳,۴۶	حمل‌ونقل ریلی	۱,۴۴
ساخت محصولات از لاستیک و پلاستیک	۳,۴۵	جنگلداری	۱,۴۳
ساخت مبلمان و مصنوعات طبقه‌بندی نشده در جای دیگر	۳,۴۲		

مأخذ: یافته‌های محقق.

جدول ۱ نتایج حاصل از الگوریتم FastPCS است. شاخص دورافتادگی[1] برای هر بخش محاسبه‌شده که بیشترین مقدار برای بخش نفت خام و گاز طبیعی است. این نشان می‌دهد که این بخش فاصله زیادی با سایر بخش‌ها دارد. دو بخش بعدی نیز (ساخت کک، فرآورده‌های نفتی و ...، ساخت چوب و محصولات چوبی) فاصله زیادی از ابرداده‌ها دارند. با توجه به سطح آستانه FastPCS، ۲۰ بخش مشخص‌شده در جدول ۱ به‌عنوان داده پرت شناسایی شدند.

با استفاده از شناسایی داده‌های پرت، داده‌های در دست بررسی، به دو گروه داده‌های پرت و بخش‌های باقیمانده تقسیم شدند. این دو گروه به‌صورت جداگانه با استفاده از الگوریتم خوشه‌بندی فازی، خوشه‌بندی می‌شوند؛ بنابراین ماتریس عدم تشابه (فاصله) با استفاده از معادله (۴,۳) برای هر دو گروه به‌طور جداگانه محاسبه می‌شود.

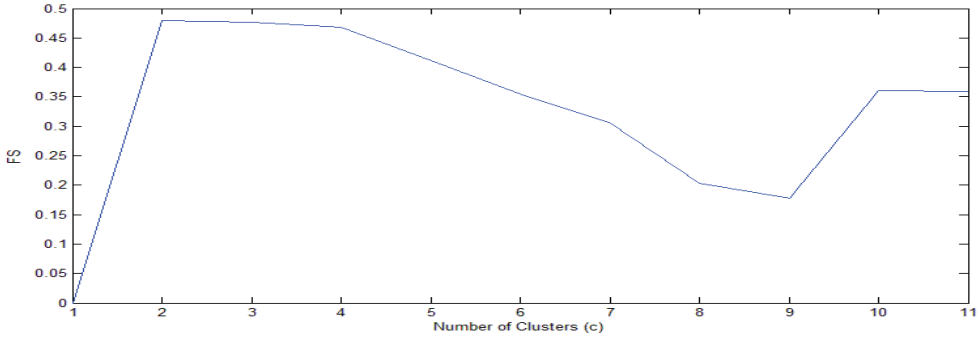
۵- نتایج تجربی

خوشه‌بندی فازی گروه داده‌های پرت

خوشه‌بندی گروه داده‌های پرت (۲۰ بخش مشخص‌شده در بخش داده‌ها)، با استفاده از الگوریتم فازی NERFCM صورت گرفت. پیش از آن، با مشاهده داده‌ها مشخص شد بخش نفت خام و گاز طبیعی فاصله زیادی با سایر نقاط این گروه دارد و به‌صورت دستی، به‌منظور بهتر خوشه‌بندی شدن سایر بخش‌ها، در خوشه جداگانه قرار داده شده است.[2] پس‌ازآن، ۱۹ بخش دیگر، برای تعیین تعداد خوشه مناسب، با الگوریتم خوشه‌بندی برای تعداد خوشه (C) از ۲ تا ۱۱ خوشه‌بندی و معیار سیلووت فازی (FS) محاسبه شد.

1- Outlyingness Index

۲- با اجرای الگوریتم خوشه‌بندی فازی برای گروه داده‌های پرت به همراه بخش «نفت خام و گاز طبیعی» مشخص شد که این بخش با درجه عضویت ۱ به‌تنهایی در یک گروه قرار دارد و فاصله زیادی با سایر بخش‌ها دارد.



شکل ۱- معیار سیلووت فازی به ازای مقادیر مختلف c (گروه داده‌های پرت)

مأخذ: یافته‌های محقق.

شکل ۱ نمودار معیار سیلووت فازی را به ازای تعداد خوشه نشان می‌دهد. اگرچه مقدار FS برای تعداد خوشه (۲) حداکثر می‌شود، مقدار این معیار برای تعداد خوشه (۳) و (۴) نیز نزدیک به مقدار حداکثر است و برای دستیابی به تفکیک بهتر بخش‌ها از تعداد خوشه (۴) برای خوشه‌بندی استفاده می‌شود. نتایج خوشه‌بندی بخش‌های این گروه به‌صورت زیر به دست آمد:

جدول ۲- نتایج خوشه‌بندی فازی گروه داده‌های پرت

خوشه	عضو	درجه عضویت	خوشه مقابل	درجه عضویت
اول	ساخت سایر محصولات کانی غیرفلزی	۰,۸۱	چهارم	۰,۱۵
	ساخت فلزات اساسی	۰,۵۰	چهارم	۰,۲۶
	ساخت ماشین‌آلات و تجهیزات طبقه‌بندی نشده در جای دیگر	۰,۵۰	چهارم	۰,۲۷
	توزیع گاز طبیعی	۰,۵۰	چهارم	۰,۲۵
	سایر معادن	۰,۴۴	دوم	۰,۲۷
	ساخت کک، فرآورده‌های حاصل از تصفیه نفت و سوخت‌های هسته‌ای	۰,۳۴	چهارم	۰,۳۲
دوم	خمیر کاغذ، کاغذ و محصولات کاغذی، اوراق چاپی و کالاهای مربوط	۰,۹۷	اول	۰,۰۱

ادامه جدول ۲

خوشه	عضو	درجه عضویت	خوشه مقابل	درجه عضویت
دوم	ساخت چوب و محصولات چوبی	۰,۸۲	اول	۰,۰۷
سوم	حمل و نقل جاده‌ای	۰,۷۶	چهارم	۰,۱۳
	زراعت و باغداری	۰,۶۵	چهارم	۰,۱۵
	عمده‌فروشی، خرده‌فروشی، تعمیر وسایل نقلیه و کالاها	۰,۶۴	چهارم	۰,۱۷
	امور عمومی	۰,۵۹	چهارم	۰,۲۴
چهارم	سایرمحصولات‌غذایی و آشامیدنی و محصولات از توتون و تنباکو	۰,۷۷	اول	۰,۱۳
	ساخت وسایل نقلیه موتوری، تریلر و نیم تریلر	۰,۶۴	اول	۰,۲۱
	هتل و خوابگاه و رستوران	۰,۴۳	اول	۰,۳۵
	حمل‌ونقل آبی	۰,۴۲	اول	۰,۳۴
	ساخت منسوجات	۰,۳۹	اول	۰,۲۵
	خدمات مستغلات	۰,۳۴	سوم	۰,۳۰
	ساخت رادیو و تلویزیون، دستگاه‌ها و وسایل ارتباطی	۰,۳۰	دوم	۰,۲۹
پنجم	نفت خام و گاز طبیعی	۱	--	--

مأخذ: یافته‌های محقق.

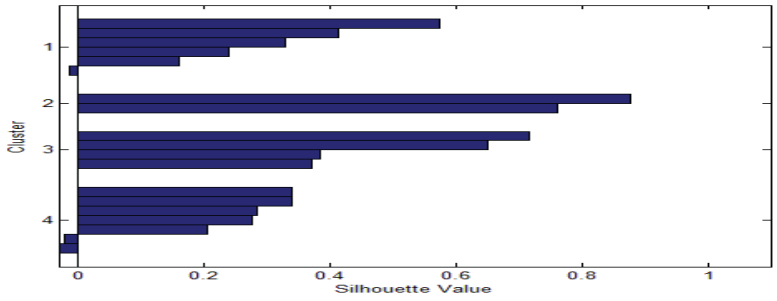
جدول ۲ خوشه‌بندی فازی گروه داده‌های پرت را به همراه درجه عضویت در خوشه، همچنین نزدیک‌ترین خوشه به هر بخش (خوشه مقابل)[1] و درجه عضویت مربوط نشان می‌دهد. با توجه به درجات عضویت، خوشه‌های دوم، سوم و پنجم به‌خوبی از هم تفکیک شده‌اند. در خوشه اول، ۵ بخش اول به‌درستی در خوشه خود قرارگرفته است؛ اما آخرین بخش، «ساخت کک، فرآورده‌های حاصل از تصفیه نفت و سوخت‌های هسته‌ای» با درجه

1- Alternative Cluster

عضویت پایین در خوشه خود قرار دارد و به درجه عضویت در خوشه مقابل آن (خوشه چهارم) نزدیک است. همچنین شکل ۲ نشان می‌دهد که خوشه‌بندی این بخش با مشکل مواجه بوده است. در خوشه چهارم نیز پنج بخش اول به‌درستی در خوشه خود قرارگرفته‌اند؛ اما دو بخش آخر (خدمات مستغلات و ساخت رادیو و تلویزیون، دستگاه‌ها و وسایل ارتباطی) با توجه به درجات عضویت و مقادیر فاصله سیلوتشان، با مشکل خوشه‌بندی مواجه بوده‌اند و در مرز خوشه‌ها قرار دارند (بخش خدمات مستغلات با توجه به درجه عضویت در خوشه اصلی و خوشه مقابل خود، بین دو خوشه چهارم و سوم، و بخش ساخت رادیو و تلویزیون، دستگاه‌ها و وسایل ارتباطی بین دو خوشه چهارم و دوم قرار دارند).

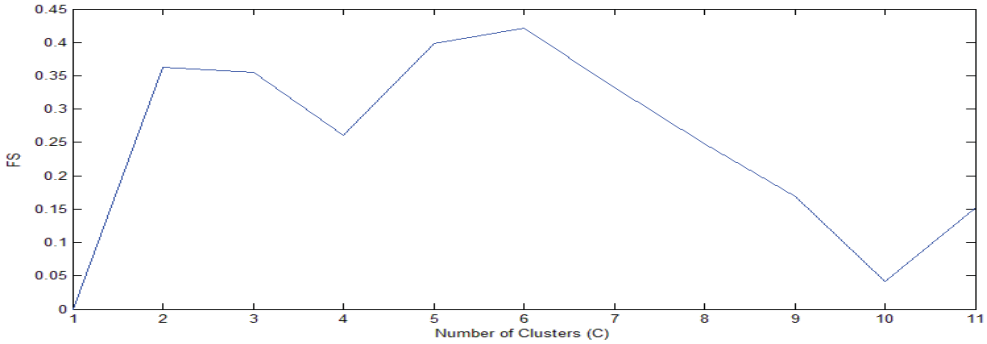
خوشه‌بندی فازی گروه باقیمانده

گروه بخش‌های باقیمانده (۲۷ بخش باقیمانده از شناسایی داده‌های پرت) مشابه گروه قبل با استفاده از الگوریتم NERFCM خوشه‌بندی شد. ابتدا برای یافتن بهترین تعداد خوشه، الگوریتم برای تعداد خوشه (C) از (۲) تا (۱۱) اجرا و معیار سیلووت فازی برای هر C محاسبه شد. شکل ۳ نشان می‌دهد که تعداد خوشه مناسب برای خوشه‌بندی (۶) خوشه است.



شکل ۲- معیار فاصله سیلووت برای بخش‌های گروه داده‌های پرت

مأخذ: یافته‌های محقق.



شکل ۳- معیار سیلووت فازی به ازای مقادیر مختلف c

مأخذ: یافته‌های محقق.

پس از تعیین تعداد خوشه مناسب، بخش‌های این گروه در ۶ خوشه، خوشه‌بندی شدند که نتایج زیر به دست آمد:

جدول ۳- نتایج خوشه‌بندی گروه بخش‌های باقیمانده

خوشه	عضو	درجه عضویت	خوشه مقابل	درجه عضویت
اول	ساختمان‌های مسکونی	۰,۹۷	چهارم	۰,۰۱
	سایر ساختمان‌ها	۰,۷۲	ششم	۰,۰۸
	خدمات آموزش عمومی- فنی‌حرفه‌ای	۰,۴۴	چهارم	۰,۱۵
	دامداری، مرغداری، پرورش کرم ابریشم و زنبورعسل و شکار	۰,۳۸	دوم	۰,۱۷
دوم	ساخت ابزار پزشکی، ابزار اپتیکی، ابزار دقیق و انواع ساعت	۰,۷۴	پنجم	۰,۰۸
	خدمات پشتیبانی و انبارداری	۰,۷۲	چهارم	۰,۱
	حمل‌ونقل ریلی	۰,۴۴	چهارم	۰,۲۹
	برق	۰,۳۷	پنجم	۰,۲۵
	بیمه	۰,۲۸	ششم	۰,۱۹
	ساخت ماشین‌آلات دفتری، حسابداری و محاسباتی	۰,۲۵	پنجم	۰,۲۲

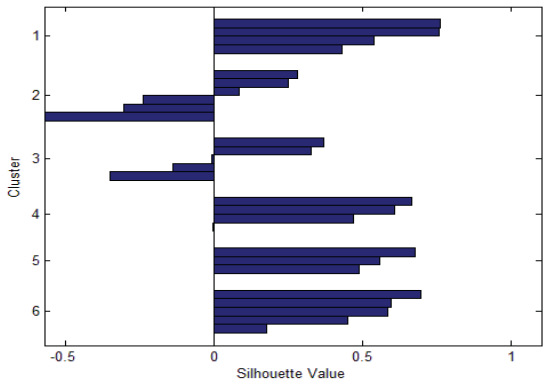
ادامه جدول ۳

خوشه	عضو	درجه عضویت	خوشه مقابل	درجه عضویت
سوم	ساخت پوشاک، عمل‌آوری و رنگ کردن خز	۰,۷۹	چهارم	۰,۰۶
	دباغی و پرداخت چرم و سایر محصولات چرمی	۰,۵۸	چهارم	۰,۱۷
	ساخت مبلمان و مصنوعات طبقه‌بندی نشده در جای دیگر	۰,۴۵	چهارم	۰,۱۶
	سایر خدمات	۰,۳۴	چهارم	۰,۱۷
	آب	۰,۲۴	پنجم	۰,۲۲
چهارم	حمل‌ونقل هوایی	۰,۸۷	دوم	۰,۰۴
	حمل‌ونقل لوله‌ای	۰,۴۶	دوم	۰,۱۴
	جنگلداری	۰,۳۵	سوم	۰,۲۱
	ماهیگیری	۰,۳۱	دوم	۰,۱۸
پنجم	ساخت محصولات فلزی فابریکی به‌جز ماشین‌آلات و تجهیزات	۰,۸۷	دوم	۰,۰۵
	ساخت محصولات از لاستیک و پلاستیک	۰,۴۲	دوم	۰,۱۹
	ساخت ماشین‌آلات و دستگاه‌های برقی طبقه‌بندی نشده در جای دیگر	۰,۳۹	چهارم	۰,۱۷
ششم	آموزش عالی	۰,۷۹	دوم	۰,۰۶
	بهداشت و درمان	۰,۶۲	چهارم	۰,۰۷
	تفریحی، فرهنگی، و ورزشی	۰,۵۶	چهارم	۰,۱۵
	پست و مخابرات	۰,۳۷	سوم	۰,۱۵
	بانک و واسطه‌گری مالی	۰,۳۴	پنجم	۰,۱۷

مأخذ: یافته‌های محقق.

خوشه‌های اول، چهارم، پنجم و ششم با توجه به مقدار فاصله سیلووت در شکل ۴ به‌درستی خوشه‌بندی شده‌اند؛ اما سه بخش آخر در دو خوشه دوم و سوم با مشکل مواجه شده است و به‌نوعی در مرز خوشه‌ها قرار دارد. در خوشه دوم، بخش «برق» و «ساخت ماشین‌آلات دفتری، حسابداری و محاسباتی» بین دو خوشه دوم و پنجم قرار دارد و بخش «بیمه» با توجه به درجه عضویت در خوشه‌های مختلف بین خوشه‌های دوم، ششم و چهارم قرار دارد.

سه بخش آخر خوشه سوم نیز با مشکل خوشه‌بندی همراه است. بخش «ساخت مبلمان و مصنوعات طبقه‌بندی نشده در جای دیگر» با درجه عضویت ۰,۴۵ در خوشه سوم و ۰,۱۶ در خوشه چهارم، در خوشه سوم و در نزدیکی خوشه چهارم قرارگرفته است. بخش «سایر خدمات» با درجه عضویت ۰,۳۴ در خوشه سوم و درجه عضویت خوشه مقابل ۰,۱۷، در خوشه سوم و نزدیک به خوشه چهارم؛ اما بخش آب به‌سختی در این خوشه قرارگرفته است(درجه عضویت در خوشه سوم ۰,۲۴) و با توجه به درجات عضویت در خوشه‌های دیگر در خوشه خاصی قرار نمی‌گیرد.



شکل ۴- معیار فاصله سیلووت برای بخش‌های گروه باقیمانده

۶- تفسیر نتایج خوشه‌بندی

پس از خوشه‌بندی بخش‌ها در دو گروه مختلف (گروه داده‌های پرت و باقیمانده) به بررسی شاخص‌های تعریف‌شده برای تعیین بخش‌های کلیدی پرداخته می‌شود. بیش از ۷۰

درصد تولید و ۹۵ درصد صادرات در بخش‌هایی صورت گرفته که به‌عنوان داده پرت شناسایی شده‌اند. همچنین سهم اشتغالی که این بخش‌ها در اشتغال کشور داشته‌اند بیش از ۶۰ درصد است؛ بنابراین می‌توان این انتظار را داشت که بخش‌های کلیدی، در گروه داده‌های پرت وجود داشته باشند.

جدول ۴- نسبت تولید، صادرات، اشتغال هر گروه به کل تولید، صادرات و اشتغال

گروه	تولید	صادرات	اشتغال
داده‌های پرت (۲۰ بخش)	٪۷۱,۶	٪۹۶,۶۸	٪۶۱,۱۷
باقیمانده (۲۷ بخش)	٪۲۸,۳۲	٪۳,۳۲	٪۳۸,۸۳

مأخذ: جدول داده- ستانده ۱۳۹۰، سرشماری نفوس و مسکن ۱۳۹۰ و یافته‌های محقق.

جدول ۵ نسبت مجموع تولید، صادرات و اشتغال هر خوشه به کل اقتصاد ایران را نشان می‌دهد. این نسبت بیانگر این است که چه میزان از تولید، صادرات و اشتغال را مجموع بخش‌های موجود در هر خوشه ایجاد می‌کند.

جدول ۵- نسبت تولید، صادرات و اشتغال هر خوشه به کل تولید، صادرات و اشتغال

خوشه	تولید	صادرات	اشتغال
اول	٪۱۹,۵	٪۲۰,۲	٪۴,۶۷
دوم	٪۰,۴۰	٪۰,۰۵	٪۱,۲۳
سوم	٪۲۳,۲	٪۹,۳۹	٪۴۵,۹
چهارم	٪۱۸,۶	٪۵,۴۴	٪۸,۸۴
پنجم	٪۱۰,۴	٪۶۱,۷	٪۰,۴۸

مأخذ: جدول داده- ستانده ۱۳۹۰، سرشماری نفوس و مسکن ۱۳۹۰ و یافته‌های محقق.

جدول ۶ نیز نسبت میانگین هر خوشه به میانگین کل برای شاخص‌های منتخب است (این نسبت در عدد ۱۰۰ ضرب شده است).

جدول ۶- نسبت میانگین هر خوشه به میانگین کل هر شاخص (گروه داده‌های پرت)

شاخص	خوشه ۱	خوشه ۲	خوشه ۳	خوشه ۴	خوشه ۵
تولید	۱۵۲	۹	۲۷۳	۱۲۵	۴۸۹
صادرات	۱۵۸	۱	۱۱۰	۳۷	۲۸۹۸
اشتغال	۳۷	۲۹	۵۴۰	۵۹	۲۳
رشد تولید	۶۲۵	-۱۸۷	۶۶	-۹۵	-۶۹
ضریب گش	۱۴۱	۳۵۶	۶۵	۹۷	۵۴
معکوس پراکندگی ضریب گش	۱۳۰	۱۹۴	۸۰	۸۷	۷۰

مأخذ: یافته‌های محقق.

با توجه به جدول ۴ و جدول ۵، وضعیت خوشه دوم در شاخص‌های تولید، صادرات و اشتغال بسیار ضعیف است و تنها در شاخص‌های بین‌بخشی در مقایسه با سایر خوشه‌ها وضعیت بهتری دارد؛ بنابراین، می‌توان این خوشه را از جمع بخش‌های کلیدی کنار گذاشت. خوشه چهارم نیز به‌جز شاخص تولید (در جدول ۶)، در سایر شاخص‌ها از میانگین کل کمتر است. خوشه سوم که از سه بخش خدمات (حمل‌ونقل جاده‌ای، امور عمومی، عمده‌فروشی، خرده‌فروشی، تعمیر وسایل نقلیه و کالاها) و یک بخش «زراعت و باغداری» تشکیل‌شده، در شاخص اشتغال، بیشترین مقدار را در بین خوشه‌ها دارد و از نظر شاخص تولید در بین خوشه‌ها پس از خوشه پنجم، دوم است. این چهار بخش، به‌تنهایی ۴۶ درصد اشتغال اقتصاد ایران را ایجاد کرده است؛ بنابراین از منظر تولید و اشتغال این چهار بخش را می‌توان به‌عنوان بخش کلیدی در نظر گرفت.

بخش نفت خام و گاز طبیعی (خوشه پنجم) در شاخص‌های تولید و صادرات در بین سایر خوشه‌ها بیشترین مقدار را دارد؛ اما ازنظر شاخص‌های بین‌بخشی که رابطه بخش با سایر بخش‌های اقتصادی را نشان می‌دهد، در مقایسه با میانگین کل مقادیر پایینی دارد و این نشان می‌دهد که این بخش، ارتباط کمی با سایر بخش‌های اقتصادی دارد. این بخش، از منظر تولید و صادرات می‌تواند بخش کلیدی تلقی شود.

خوشه اول، که خوشه بخش‌های صنعتی و معدنی است، ازنظر شاخص‌های تعریف‌شده، وضعیت مطلوبی دارد. تمامی شاخص‌ها به‌جز اشتغال، از میانگین کل بیشتر هستند و میانگین رشد تولید این بخش‌ها در مقایسه با میانگین کل بسیار بالا بوده است. ازنظر شاخص‌های بین‌بخشی نیز پس از خوشه دوم، بیشترین مقدار را دارد و در شاخص صادرات به‌جز خوشه پنجم، در مقایسه با سایر خوشه‌ها مقدار بیشتری دارد. درمجموع، بخش‌های موجود در این خوشه، ۲۰ درصد از تولید کل اقتصاد و ۲۰ درصد صادرات اقتصاد کشور را ایجاد می‌کنند و با توجه به شاخص‌های بین‌بخشی، روابط مناسب و مطلوبی با سایر بخش‌های اقتصاد دارند. این بخش‌ها را می‌توان به‌عنوان بخش‌های کلیدی در ایجاد تحرک اقتصادی در سایر بخش‌ها دانست و به‌منظور رسیدن به رشد حداکثر، می‌توان در این بخش‌ها سرمایه‌گذاری کرد.

با توجه به جدول ۷، در بین شش خوشه گروه بخش‌های باقیمانده، ازلحاظ شاخص‌های انتخاب‌شده، هیچ‌کدام از خوشه‌ها شرایط بخش کلیدی را ندارند. همان‌طور که گفته شد، درمجموع، سهم تولید این ۲۷ بخش در اقتصاد ۲۸ درصد و سهم صادراتشان کمتر از ۴ درصد است. خوشه اول در این گروه تنها در دو شاخص تولید و اشتغال از میانگین کل بیشتر است و در سایر شاخص‌ها وضعیت مطلوبی ندارد. خوشه دوم نیز از نظر شاخص‌های تولید، صادرات و اشتغال در حد بسیار پایینی قرار دارد. خوشه سوم و چهارم و ششم نیز در تمامی شاخص‌ها از میانگین کل پایین‌تر است (رشد تولید خوشه ششم از میانگین کل بالاتر است، اما با توجه به پایین بودن تولید اهمیت چندانی ندارد). خوشه پنجم به‌جز در شاخص‌های بین‌بخشی در سایر شاخص‌ها از میانگین کل پایین‌تر است.

جدول ۷- نسبت میانگین هر خوشه به میانگین کل هر شاخص (گروه بخش‌های باقیمانده)

شاخص	خوشه ۱	خوشه ۲	خوشه ۳	خوشه ۴	خوشه ۵	خوشه ۶
تولید	۱۵۲	۲۰	۱۷	۷	۳۵	۷۳
صادرات	۱	۶	۸	۶	۱۵	۱
اشتغال	۲۵۹	۱۳	۴۶	۶	۴۲	۶۶
رشد تولید	۱۰۰	۳۷۵	-۳۸۸	۷	۹۷	۱۷۶
ضریب گش	۵۷	۹۴	۷۵	۷۱	۱۳۴	۶۰
معکوس پراکندگی ضریب گش	۷۰	۱۰۹	۹۳	۸۳	۱۴۷	۷۲

مأخذ: یافته‌های محقق.

۷- خلاصه و جمع‌بندی

نتایج به‌دست‌آمده از اجرای مدل و تفسیر آن نشان می‌دهد بخش‌هایی که به‌عنوان بخش کلیدی انتخاب‌شده‌اند، در گروه داده‌های پرت قرار دارند. این نتیجه بیانگر آن است که شناسایی و جداسازی داده‌های پرت (همان‌طور که موریلیاس و دیاز (۲۰۰۸) در مقاله خود اشاره می‌کنند) روشی کاربردی برای شناسایی بخش‌های کلیدی اقتصاد است. همچنین استفاده از خوشه‌بندی فازی به‌جای خوشه‌بندی سخت، درک بهتری از میزان تناسب و تشابه بخش‌ها به یکدیگر را ارائه می‌دهد؛ به‌طور مثال، بخش «ساخت کک، فرآورده‌های حاصل از تصفیه نفت و سوخت‌های هسته‌ای» که در خوشه اول از گروه داده‌های پرت قرار دارد درجه عضویت پایینی در مقایسه با سایر اعضای خوشه خود است (درجه عضوت این بخش در خوشه خود ۰,۳۴ است) این نشان می‌دهد که این بخش در مقایسه با هم‌گروهان خود تشابه کمتری دارد.

نتیجه دیگر به‌دست‌آمده از این پژوهش، بهبود فنون شناسایی داده- ستانده با ترکیب کردن این روش‌ها با خوشه‌بندی و استفاده از شاخص‌های دیگر ازجمله تولید، صادرات، اشتغال و... هر بخش است؛ برای مثال در بیشتر پژوهش‌های انجام‌شده در زمینه شناسایی

بخش‌های کلیدی اقتصاد ایران، بخش‌های «خمیرکاغذ، کاغذ و محصولات کاغذی، اوراق چاپی و کالاهای مربوط» و «ساخت چوب و محصولات چوبی» بخش‌های کلیدی اقتصاد ایران هستند؛ اما شاخص‌های تولید، صادرات و اشتغال این بخش‌ها نشان می‌دهد وزن پایینی در اقتصاد ایران دارند. همچنین در دوره ۵ ساله ۱۳۸۵ تا ۱۳۹۰ رشدی منفی در تولید و صادرات داشته‌اند، اگرچه از لحاظ شاخص‌های بین‌بخشی مقادیر بالایی دارند، نمی‌توانند بخش‌های کلیدی اقتصاد ایران باشند.

درنهایت، نتایج خوشه‌بندی نشان می‌دهد بخش‌های کلیدی اقتصاد ایران، عمدتاً بخش‌های صنعتی و معدنی هستند. خوشه اول در گروه داده‌های پرت، در تمامی شاخص‌های تعریف‌شده، به‌طور نسبی شرایط مطلوبی در مقایسه با سایر خوشه‌ها- چه در گروه داده‌های پرت و چه در گروه بخش‌های باقیمانده- است. این خوشه که از بخش‌های «ساخت سایر محصولات کانی غیر فلزی»، «ساخت فلزات اساسی»، «ساخت ماشین‌آلات و تجهیزات طبقه‌بندی نشده در جای دیگر»، «توزیع گاز طبیعی»، «سایر معادن» و «ساخت کک، فرآورده‌های حاصل از تصفیه نفت و سوخت‌های هسته‌ای» تشکیل‌شده است، درمجموع نزدیک به ۲۰ درصد از تولید اقتصاد ایران در سال ۱۳۹۰ را ایجاد کرده است و همچنین ۲۰ درصد صادرات ایران را این خوشه صورت داده است؛ بنابراین می‌توان این بخش‌ها را بخش‌های کلیدی اقتصاد ایران در نظر گرفت.

همچنین با توجه به شاخص‌های موجود و نتایج خوشه‌بندی، می‌توان بخش «نفت خام و گاز طبیعی» را از منظر صادرات، همچنین بخش‌های «عمده‌فروشی، خرده فروشی، تعمیر وسایل نقلیه و کالاها»، «امور عمومی»، «حمل‌ونقل جاده‌ای» و «کشاورزی و باغداری» را که در خوشه سوم گروه داده‌های پرت قرار دارند از منظر اشتغال‌زایی بخش‌های مهم اقتصاد ایران دانست.

منابع

بانک مرکزی جمهوری اسلامی ایران (۱۳۸۳)، آمار سری زمانی حساب‌های ملی به قیمت جاری و ثابت.

جهانگرد، اسفندیار (۱۳۹۳)، *تحلیل‌های داده- ستانده، فناوری، برنامه‌ریزی و توسعه*، تهران: نشر آماره.

جهانگرد، اسفندیار و نیلوفرالسادات حسینی (۱۳۹۲)، «شناسایی بخش‌های کلیدی اقتصاد ایران بر مبنای تحلیل تصادفی داده- ستانده (SIO)»، *مجله تحقیقات مدل‌سازی اقتصادی* (۱۱): ۴۸- ۲۳.

جهانگرد، اسفندیار و پردیس عاشوری (۱۳۸۹)، «شناسایی بخش‌های کلیدی با رویکردهای تحلیل داده- ستانده (IO)، اقتصادسنجی (EC) و تحلیل پوششی داده‌ها (DEA): مطالعه موردی ایران»، *مجله سیاست‌گذاری اقتصادی* (۳): ۱۳۵- ۱۰۷.

جهانگرد، اسفندیار و ویدا کشت‌ورز (۱۳۹۰) «شناسایی بخش‌های کلیدی اقتصاد ایران: رویکرد نوین نظریه‌ی شبکه»، *مجله اقتصاد و تجارت نوین* (۲۵ و ۲۶): ۱۲۰- ۹۷.

فنی ممتاز، هادی (۱۳۹۰) «شناسایی بخش‌های کلیدی اقتصاد ایران: رویکرد تلفیقی داده- ستانده و فازی»، *پایان‌نامه کارشناسی ارشد*، دانشگاه علامه طباطبایی، شماره بازیابی (۹۹۹۶پ).

قره‌باغبان، مرتضی (۱۳۸۷)، *اقتصاد رشد و توسعه*، تهران: انتشارات جهاد دانشگاهی.

کلانتری، خلیل (۱۳۸۰)، *برنامه‌ریزی و توسعه منطقه‌ای (تئوری‌ها و تکنیک‌ها)*، تهران: انتشارات خوشبین و انوار دانش.

گتاک، سابرتا (۱۳۶۹)، *اقتصاد توسعه*، ترجمه زهرا افشاری، تهران: انتشارات جهاد دانشگاهی.

مرکز آمار ایران، سرشماری نفوس و مسکن ۱۳۹۰.

مرکز آمار ایران، سرشماری نفوس و مسکن ۱۳۸۵.

مرکز پژوهش‌های مجلس شورای اسلامی (۱۳۹۳) جدول داده- ستانده سال ۱۳۹۰ اقتصاد ایران.

مرکز پژوهش‌های مجلس شورای اسلامی (۱۳۹۱) جدول داده- ستانده سال ۱۳۸۵ اقتصاد ایران.

مؤمنی، منصور (۱۳۹۰)، *خوشه‌بندی داده‌ها* (تحلیل خوشه‌ای)، تهران، ناشر: مؤلف.

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). "FCM: The Fuzzy c-means Clustering Algorithm". *Computers & Geosciences*, 10(2), 191-203.

Campello, R. J., & Hruschka, E. R. (2006). "A Fuzzy Extension of the Silhouette width Criterion for Cluster Analysis". *Fuzzy Sets and Systems*, 157(21), 2858-2875.

Díaz, B., Moniche, L., & Morillas, A. (2006). "A Fuzzy Clustering Approach to the Key Sectors of the Spanish Economy". *Economic Systems Research*, 18(3), 299-318.

Hathaway, R. J., & Bezdek, J. C. (1994). "NERF c-means: Non-Euclidean Relational Fuzzy Clustering". *Pattern Recognition*, 27(3), 429-437.

Huang, Z., & Michael K. Ng (1999). "A Fuzzy k-modes Algorithm for Clustering Categorical Data. Fuzzy Systems", IEEE Transactions on, 7(4), 446-452.

Morillas, A., & Diaz, B. (2008). "Key Sectors, Industrial Clustering and Multivariate Outliers". *Economic Systems Research*, 20(1), 57-73.

Vakili, K., & Schmitt, E. (2014). "Finding Multivariate Outliers with FastPCS". *Computational Statistics & Data Analysis*, 69, 54-66.

Zaki, M. J., & Meira Jr, W. (2014). "Data Mining and Analysis: Fundamental Concepts and Algorithms". *Cambridge University Press.*