مدلسازی داده‌های مرکب خاک (اجزاء بافت خاک) با استفاده از رگرسیون دریچلت

علیرضا امیریان چکان[۱]، زهرا درویش پسند[۲]، رخسار اکبری فضلی[۳]، صاحب خرده‌بین[۱]، روح‌الله تقی‌زاده مهرجردی[۴]
۱- به ترتیب استادیار و کارشناس ارشد دانشگاه صنعتی خاتم‌الانبیاء بهبهان ۲- کارشناس ارشد علوم خاک ۳- دانشجوی دکتری علوم خاک، دانشگاه آزاد اسلامی، واحد علوم و تحقیقات خوزستان ۴- استادیار دانشگاه اردکان

چکیده

داده‌های مرکب نسبتهای (درصدهای) غیر منفی هستند که مجموع آنها برابر با یک عدد ثابت است (مثل بافت خاک). به دلیل قید ثابت بودن جمع مقادیر تخمینی و وجود همبستگی بین آنها، مفاهیم کوواریانس و همبستگی کلاسیک برای این داده‌ها اغلب گمراه کننده و نامناسب است. اگر چه تبدیل لگاریتمی نسبت اجزاء داده مرکب رویکردی است که شرط ثابت بودن مجموع نسبتها پس از تخمین را تضمین می‌کند ولی چون پارامترهای به دست آمده داده اصلی نیستند، تفسیر آنها مشکل است. در این بررسی رگرسیون دریچلت که روشی انعطاف‌پذیر برای توصیف ساختارهای مختلف کوواریانس در داده‌های مرکب است و فاقد محدودیتهای ذکر شده می‌باشد برای تخمین بافت خاک با استفاده از داده‌های کمکی مورد استفاده قرار گرفت. آزمون معنی‌داری ضرایب رگرسیون نشان داد ($P<0.001$) این مدل می‌تواند جایگزینی مناسب برای روشهای کلاسیک آماری و روشهای تبدیل لگاریتمی برای تخمین اجزاء بافت خاک باشد.

واژه‌های کلیدی: داده مرکب، توزیع دریچلت، جزء بافت خاک

مقدمه

داده‌های مرکب[۱] داده‌های غیر منفی هستند که حاوی اطلاعات نسبی هستند و مجموع عناصر آنها در هر نمونه برابر با یک عدد ثابت است (یک یا ۱۰۰). داده‌های بافت خاک رایج‌ترین داده‌های مرکب در علوم خاک هستند (Odeh et al., 2003). زیرا بافت خاک از سه جزء رس، سیلت و شن تشکیل شده است و جمع آنها برابر با ۱۰۰ درصد است. در این داده‌ها به دلیل اینکه قید برابر بودن مجموع اجزاء برابر با یک عدد ثابت وجود دارد، مفاهیم اصلی کوواریانس و همبستگی که در روشهای رایج آماری مورد بررسی قرار می‌گیرند، گمراه کننده هستند. همچنین استفاده از روشهای رایج رگرسیونی برای داده‌های مرکب اغلب منجر به نتایجی از جمله عدم همبستگی بین متغیر پاسخ می‌گردد که برای این نوع داده‌ها منطقی نیست (Wang et al., 2013). این در حالی است که حداقل بین دو جزء از اجزاء یک داده مرکب سه جزیی مثل بافت خاک همبستگی وجود دارد. بنابراین بهتر است این داده‌ها با روشهای رایج آماری تجزیه و تحلیل نشوند.

مشکل دیگر داده‌های مرکب این است که اگر اجزاء یک ترکیب جداگانه تخمین زده شوند، مجموع جزءهای تخمینی الزاما برابر با یک مقدار ثابت نمی‌شود (Lark and Bishop, 2007; Odeh et al., 2003). ایتچسون (Aitchison, 1982) تبدیل لگاریتمی نسبت اجزاء داده مرکب را برای حل این مشکلات پیشنهاد نمود. یکی از مشکلات این رویکرد این است که در این روش بیشتر تاکید روی اجزایی است که از تغییرات نسبی بیشتری برخوردارند، در حالی که در اکثر مواقع اینها اجزایی هستند که درصدها و تغییرات مطلق آنها کم است (Hijazi and Jernigan, 2009). در نتیجه چون تاکید روی نسبتهاست نه مقادیر مطلق، تاثیر اجزاء مهم و تاثیر گذار در تخمین سایر خواص مرتبط با آنها ممکن است نادیده گرفته شود.

[1] - Compositional data

برای حل مشکلات ذکر شده در تجزیه و تحلیل داده‌های ترکیبی، رویکردهای متعددی ارایه شده است. به عنوان مثال توزیع تعمیم یافته Liouville[۲] و توزیع Liouville[۳] شرطی و مدل متغیر کمکی دریچلت[۴] (Hijazi. and Jernigan, 2009) را می توان نام برد. توزیع دریچلت ابتدا توسط کونور و موسیمن (Connor and Mosimann, 1969) برای داده‌های مرکب معرفی گردید. رگرسیون دریچلت[۵] روشی مناسب و انعطاف پذیر برای توصیف روندها و ساختارهای مختلف کوواریانس در داده‌های مرکب است و محدودیتهای روشهای معمول آماری و استفاده از نسبتهای لگاریتمی را ندارد. از آنجا که بیشتر متخصصین علوم خاک با مشکلات تحلیل داده‌های مرکب به روشهای معمول آماری آشنا نیستد و در اغلب مطالعات مرکب بودن بافت و ترکیبات خاک در نظر گرفته نمی‌شود، در این تحقیق کارآیی رگرسیون دریچلت برای تخمین اجزاء بافت خاک با استفاده از متغیرهای کمکی مستخرج از تصویر ماهواره‌ای و مدل رقومی ارتفاع مورد بررسی قرار گرفت.

مواد و روشها

منطقه مورد مطالعه به وسعت حدود ۴۶۰۰ هکتار در شمال شرق شهر بهبهان در استان خوزستان واقع شده است. داده‌های استفاده شده در این تحقیق درصدهای اجزاء رس، سیلت و شن خاک در ۱۰۹ نقطه مشاهداتی بودند (۳۰ سانتی‌متر سطحی). انتخاب محلهای نمونه‌برداری بر اساس روش هایپرکیوب شرطی (Minansny and McBratney, 2006) تعیین گردید. این رویکرد یک روش نمونه‌برداری تصادفی طبقه‌بندی شده است که محل‌های نمونه‌برداری را بر اساس توزیع متغیرهای کمکی که با متغیر اصلی همبستگی دارند انتخاب می‌کند.

برای ایجاد مدل رگرسیونی مورد نظر، ابتدا همبستگی بین اجزاء بافت خاک با متغیرهای کمکی مستخرج از تصاویر ماهواره‌ای و مدل رقومی ارتفاع (DEM) از جمله ارزش باندهای تصاویر ماهواره‌ای، شیب، فاکتور شیب-طول شیب، MrVBF[۶] (Gallant and Dowling, 2003)، انحنای پروفیل و انحنای سطح مورد بررسی قرار گرفت و در نهایت متغیرهای کمکی که همبستگی معنی‌داری با اجزاء بافت خاک داشتند به عنوان ورودی مدل رگرسیونی انتخاب شدند.

برای تخمین اجزاء بافت خاک از مدل رگرسیون دریچلت استفاده گردید و نتایج با مدل رگرسیونی حداقل مربعات (OLS[۷]) مقایسه گردید. رگرسیون دریچلت را می‌توان برای داده‌های مرکب یعنی حالتی که متغیر وابسته (مثل بافت خاک) از اجزایی تشکیل شده است که جمع این اجزاء برابر با یک مقدار ثابت است، به کار برد. این مدل بر اساس روشهای حداکثر احتمال به داده‌ها برازش داده می‌شود. توزیع دریچلت که نوع تعمیم یافته توزیع بتا می‌باشد در رابطه ۱ ارایه شده است. چنانچه $X = (x_1, \dots, x_D)$ یک بردار مثبت با ابعاد $1 \times D$ دارای توزیع دریچلت با پارامترهای مثبت $(\lambda_1, \dots, \lambda_D)$ باشد، تابع چگالی آن به صورت زیر است (رابطه ۱):

$$f(X) = (\Gamma(\lambda)/\prod_{i=1}^{D}\Gamma(\lambda_i))\prod_{i=1}^{D} x_i^{(\lambda_i-1)} \qquad (۱)$$

در این رابطه قیود $\sum_{i=1}^{D} x_i = 1$ و $\lambda = \sum_{i=1}^{D}\lambda_i$ برقرار هستند.

با تعیین میزان تغییر پارامترهای توزیع دریچلت با تغیرات متغیرهای کمکی، مدل رگرسیون دریچلت به سهولت به دست می‌آید. به عبارت دیگر، برای هر متغیر کمکی s، پارامترهای توزیع دریچلت را می‌توان به صورت توابعی (با مقادیر مثبت) از s در نظر گرفت. علاوه بر توابع نمایی، توابع چند جمله‌ای نیز برای این منظور مناسب هستند. در این تحقیق رگرسیون دریچلیت و تجزیه و تحلیل‌های مرتبط آن و همچنین رسم نمودارهای مربوطه با استفاده از بسته نرم‌افزاری "DirichletReg" در محیط نرم‌افزار R انجام گرفت.

---

[2] - Generalized Liouville distribution
[3] - Conditional Liouville distribution
[4] - Dirichlet covariate model
[5] - Dirichlet regression
[6] - Multiresolution Index of Valley Bottom Flatness
[7] - Ordinary least square regression

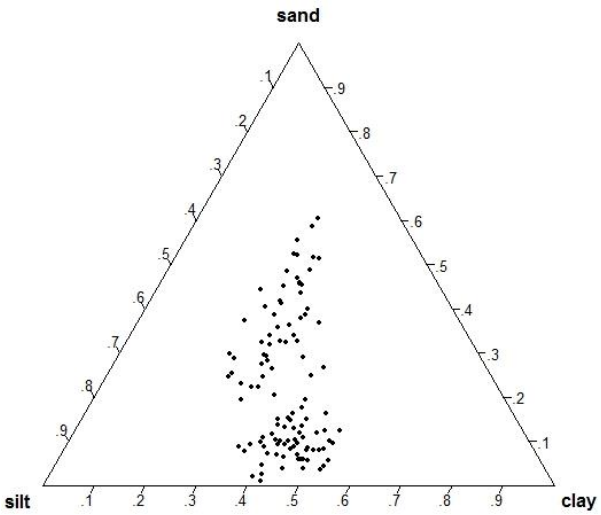
نتایج و بحث

مقادیر مرکز و واریانس[8] کل داده‌های بافت خاک در جدول ۱ ارایه شده است. برای داد ه های مرکب به جای میانگین بهتر است از مرکز آنها که تقریبا معادل میانگین هندسی است استفاده کرد. مقدار بالای واریانس کل (واریانس رس، سیلت و شن) نشان دهنده این است که داده‌ها به طور کلی از تغییرپذیری نسبتا زیادی برخوردار هستند.

جدول۱- مرکز و واریانس کل داده‌های بافت خاک

جزء بافت	مرکز (%)	واریانس کل
رس	۳۵/۵۰	
سیلت	۴۳/۳۰	۰/۷۷
شن	۱۵/۱۸	

مقادیر نسبتهای رس، سیلت و شن برای همه نقاط مطالعاتی در شکل ۱ ارایه شده است. بر اساس شکل بیشتر نمونه‌های خاک دارای سیلت متوسط تا نسبتا زیاد، رس خیلی کم تا متوسط و شن متوسط تا خیلی زیاد هستند.



شکل ۱- توزیع درصدهای رس، سیلت و شن روی مثلث بافت خاک

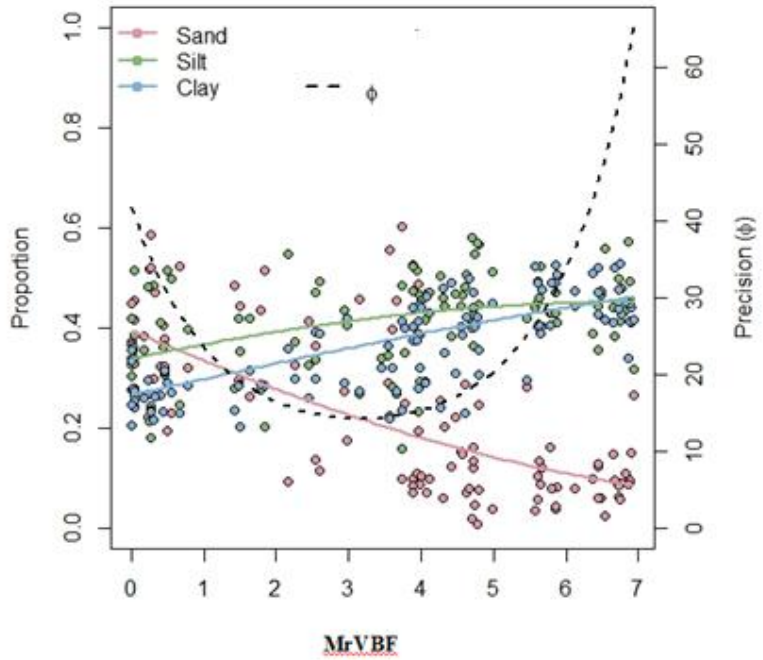
[8] - Center and total variance

با توجه به اینکه از بین متغیرهای کمکی فقط MrVBF با هر سه جزء بافت خاک همبستگی معنی داری داشت، این متغیر به عنوان متغیر مستقل برای تخمین اجزا بافت خاک مورد استفاده قرار گرفت. در جدول ۲ نتایج آزمون معنی‌داری ضرایب مدل رگرسیون دریچلت ارایه شده است. مدل رگرسیون دریچلت به دو صورت اجرا گردید؛ یک بار با MrVBF به عنوان متغیر مستقل و نسبتهای رس، سیلت و شن به عنوان متغیر وابسته و یک بار هم مجذور MrVBF به عنوان متغیر مستقل در نظر گرفته شد. نتایج نشان داد ضرایب هر دو مدل رگرسیونی در سطح ۰/۰۰۱ معنی‌دار بوده اند (جدول ۲). به عبارت دیگر ضرایب مدل رگرسیون به طور معنی‌داری مخالف صفر بودند و مدل رگرسیون دریچلت به خوبی قادر به تخمین هر سه جزء بافت خاک بوده است.

جدول ۲- ضرایب مدل رگرسیون دریچلت برای تخمین اجزاء بافت خاک

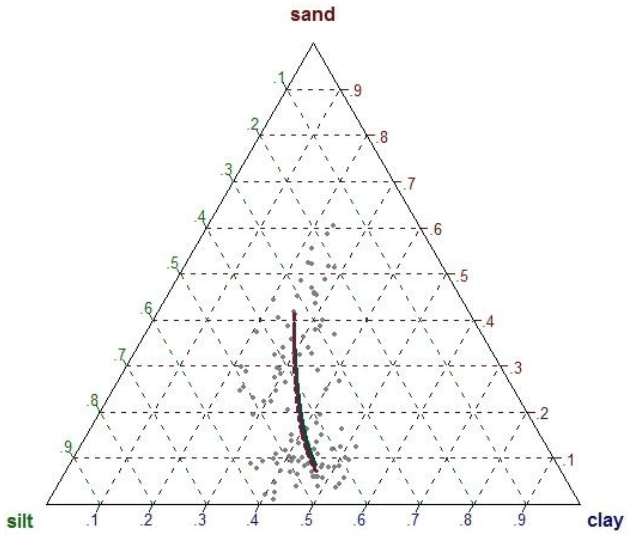
جزء بافت		ضرایب رگرسیون	خطای استاندارد	p
شن	عرض از مبدا	۲/۸۰	۰/۲۲	۰/۰۰۰***
	MrVBF	-۰/۸۳	۰/۱۵	۰/۰۰۰***
	$(MrVBF)^2$	۰/۰۹	۰/۰۲	۰/۰۰۰***
سیلت	عرض از مبدا	۲/۶۵	۰/۲۳	۰/۰۰۰***
	MrVBF	-۰/۵۹	۰/۱۵	۰/۰۰۰***
	$(MrVBF)^2$	۰/۱۰	۰/۰۲	۰/۰۰۰***
رس	عرض از مبدا	۲/۴۱	۰/۱۳	۰/۰۰۰***
	MrVBF	-۰/۵۶	۰/۱۵	۰/۰۰۰***
	$(MrVBF)^2$	۰/۱۰	۰/۰۲	۰/۰۰۰***

نتایج برازش تابع چند جمله‌ای با استفاده از مدل رگرسیون دریچلت در شکل ۲ ارایه شده است. بر اساس شکل با افزایش MrVBF مقدار شن کاهش و مقدار رس و سیلت تا حدودی افزایش می‌یابد. همچنین میزان دقت کلی مدل (ϕ) در تخمین سه جزء رس، سیلت و شن در مقادیر کم MrVBF نسبتا زیاد است و با افزایش مقدار MrVBF تا مقادیر متوسط از دقت مدل کاسته می‌شود و سپس در مقادیر بالای MrVBF دقت مدل به حداکثر می‌رسد.



شکل ۲- دقت کلی مدل رگرسیون دریچلت در تخمین اجزاء بافت خاک و رابطه بین این اجزاء با MrVBF

در شکل ۳ نتایج روش OLS (خط چین قرمز) و رگرسیون دریچلت (خط سبز) با هم مقایسه شده است. بر اساس شکل تطابق بالایی بین نتایج رگرسیون دریچلت با نتایج OLS وجو دارد.



شکل ۳- نتایج تخمینهای حاصل از رگرسیون دریچلت (خط سبز) و رگرسیون حداقل مربعات (خط قرمز)

نتایج این بررسی نشان داد که استفاده از مدل رگرسیون دریچلت رویکرد مناسبی برای تخمین داده‌های مرکب مثل بافت خاک است. در مطالعات متعددی برتری استفاده از رگرسیون دریچلت نسبت به روشهای مرسوم آماری و همچنین روشهای تبدیل نسبت داده‌ها به اثبات رسیده است (Hijazi, 2003; Hijazi, and Jernigan, 2009). بنابراین برای جلوگیری از تفسیرهای اشتباه ناشی از به کار بردن روش‌های کلاسیک آماری و یا روشهای تبدیل لگاریتمی، می‌توان از رگرسیون دریچلت در بررسی و تخمین بافت خاک در مطالعات مختلف از جمله نقشه‌برداری رقومی استفاده کرد. همچنین با توجه به اینکه از رگرسیون دریچلت

در مطالعات بافت خاک به ندرت استفاده شده است، پیشنهاد می‌شود مطالعات بیشتری در این زمینه صورت گیرد و این مدل با متغیرهای مستقل بیشتر و متنوع‌تری اجرا شود. علاوه بر بافت خاک، می‌توان ترکیب عناصر غذایی خاک و یا کانیهای خاک را نیز با استفاده از روشهای خاص تجزیه و تحلیل داده‌های مرکب مورد بررسی و مطالعه قرار داد.

منابع

Aitchison J. 1982. The statistical analysis of compositional data. Journal of the Royal Statistical Society, 44: 139-177.

Connor R. and Mosimann J. 1969. Concepts of independence for proportions with a generalization of the Dirichlet distribution. Journal of the American Statistical Association, 64: 194-206.

Gallant J.C. and Dowling T.I. 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resources Research, 39: 1347–1360.

Hijazi R.H. 2003. Analysis of compositional data using Dirichlet covariate models. PhD thesis, American University, Washington, D.C.

Hijazi R.H. and Jernigan R.W. 2009. Modeling compositional data using Dirichlet regression models. Journal of Applied Probability & Statistics, 4:77–91.

Lark R.M. and Bishop T.F.A. Cokriging particle size fractions of the soil. European Journal of Soil Science, 58: 763–774.

Minasny B. and McBratney A.B. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computer & Geosciences, 32: 1378-1388.

Odeh I.O.A., Todd A.J. and Triantafilis J. 2003. Spatial prediction of soil particle-size fraction as compositional data. Soil Science, 168: 501-515.

Wang H., Shangguan L., Wu J. and Guan R. 2013. Multiple linear regression modeling for compositional data. Neurocomputing, 122: 490–500.

Modeling soil compositional data (soil texture) using Dirichlet regression

A. Amirian Chakan[1], Z. Darvishpasand[2], R. Akbari Fazli[3], S. Khordbin[4] R. Taghizadeh-Mehrjardi[5]
1 & 4. Assistant Professor and Senior Expert, Behbahan Khatam Alanbia University of Technology
2. Senor Expert in Soil Science
3. PhD Student, Islamic Azad University, Science and Research Branch, Khuzestan
5. Assistant Professor, Ardakan University

Abstract

Compositional data are non-negative proportions (such as soil texture fractions) with a constant sum. Ensuring constant sum of the estimated elements and the correlation between them, the classical concepts of covariance and correlation are often undesirable and misleading. Although, logratio transformation guarantees constant sum of the estimated elements of any composition, but the resulting parameters have no straightforward meaning and the interpretations are difficult. Dirichlet regression (DR) which is a flexible model to explain the different covariance structures in compositional data and has none of the mentioned drawbacks, was used to estimate soil texture fractions using related covariates. Since the regression coefficients were significantly different from zero ($P<0.001$), DR can be considered as an informative alternative to classical statistical and logratio methods for estimating soil particle fractions.

Keywords: Compositional data, Dirichlet distribution, Soil texture fraction,