ESTRATIFICAÇÃO E REMOÇÃO DE OUTLIERS EM UMA BIBLIOTECA ESPECTRAL DE SOLOS, NA PREDIÇÃO DE MATÉRIA ORGÂNICA DO SOLO

Bruno dos Anjos Bartsch [1], Raul Roberto Poppiel [2],Nícolas Augusto Rosin[3], José Alexandre Melo Demattê [4]

[1]Escola Superior de Agricultura "Luiz de Queiroz" – Universidade de São Paulo, brunobartsch@usp.br;
[2]Escola Superior de Agricultura "Luiz de Queiroz" – Universidade de São Paulo, raulpoppiel@usp.br;
[3]Escola Superior de Agricultura "Luiz de Queiroz" – Universidade de São Paulo, jamdemat@usp.br;
[4]Escola Superior de Agricultura "Luiz de Queiroz" – Universidade de São Paulo, jamdemat@usp.br

RESUMO

Compreender a variabilidade da distribuição dos atributos químicos e granulométricos do solo, tais como: Areia, Argila, C.T.C e Matéria Orgânica do Solo (MOS), Macronutrientes e Micronutrientes no solo torna-se um fator fundamental no manejo a favor da qualidade do solo e na mitigação das mudanças climáticas. Uma das ferramentas necessárias para realizar esta compreensão é a análise de solo, embora as análises tradicionais por química úmida tenham sua eficiência em quantificar corretamente os teores destes atributos do solo, possui limitações, por serem métodos financeiramente custosos e exigem reagentes químicos nocivos ao meio ambiente. Nesse contexto, a análise de solo via espectroscopia, é uma estratégia a estes entraves. Desta forma, o objetivo deste trabalho é comparar o potencial de predição da MOS, via modelagem algoritmo *Cubist*, para dois tratamentos de dados espectrais, com e sem outliers, identificados em grupos de amostras espectralmente similares.

Palavras-chave **—** Matéria Orgânica, Estratificação, Outlier.

ABSTRACT

Understanding the variability of the distribution of chemical and granulometric attributes of the soil, such as: Sand, Clay, C.T.C and Soil Organic Matter (SOM), Macronutrients and Micronutrients in the soil becomes a fundamental factor in the management in favor of soil quality and in mitigating climate change. One of the tools necessary to carry out this understanding is soil analysis, although traditional analysis by wet chemistry has its efficiency in correctly quantifying the levels of these soil attributes, it has limitations, as they are financially expensive methods and require chemical reagents harmful to the environment. In this context, a soil analysis via spectroscopy is a strategy to these inputs. Thus, the objective of this work is to compare the SOM prediction potential, via Cubist algorithm modeling, for two treatments of spectral data, with and without outliers, identified in groups of spectrally similar samples.

Key words **—** *Organic Matter, Stratification, Outlier*

1. INTRODUÇÃO

O solo é fundamental no ciclo do carbono terrestre, com potencial de sequestro de 0.90 a 1.85 GtC ano^{-1} nas áreas agrícolas mundiais [13]. Uma das variáveis reconhecida como um dos principais indicadores da qualidade do solo, corresponde à matéria orgânica do solo (MOS), devido a seu efeito benéfico nas propriedades físicas e no fornecimento de nutrientes.. Diante de cenários futuros de influências antrópicas e mudanças climáticas globais, a quantificação do teor de carbono orgânico do solo (COS) é fundamental para identificar áreas onde os solos serão fontes ou sumidouros de COS. Fica evidenciado a necessidade de métodos ágeis e mais sustentáveis para quantificar o conteúdo COS em comparação com os métodos tradicionais, que são extensivos e nocivos ao meio ambiente [10], além de que os resultados dos laboratórios tradicionais estão suscetíveis a erros sistemáticos e aleatórios que frequentemente dependem deles mesmos e/ou das técnicas analíticas utilizadas, levando a *outliers* [8]. Dado o contexto, uma estratégia é o uso de técnicas de espectroscopia Vis-NIR-SWIR, a qual a amostra de solo é submetida a uma técnica não destrutiva, de baixo custo e permitindo a obtenção de um grande volume de informações do solo [10] sem gerar resíduos químicos para a predição do COS.

Para construir modelos preditivos do conteúdo de COS em diferentes escalas espaciais, os dados espectrais são organizados em Bibliotecas Espectrais do Solo (SSL). Logo o conteúdo do COS pode ser predito por modelos calibrados com SSL, a nível propriedade [7] , regional [2], nacional [9] e global [10]. Porém desenvolver modelos preditivos precisos, continua sendo um trabalho desafiador [6], em particular na escala a nível propriedade e regional, situações onde são necessários dados como maior grau de detalhamento e precisão [2]. O desafio dos modelos espectrais em predizer o conteúdo do COS, persiste quando a SSL é composta por amostras de amostras de solo formadas sob diferentes materiais de origem, relevo, clima e diferentes usos da terra. Essa variabilidade da origem do solo, corrobora para a variação do teor de COS numa dada paisagem. O efeito desta variabilidade, é a variância no SSL que pode ser grande [2], prejudicando assim o desempenho da predição dos modelos [10]. Outro fator que pode

corroborar para o aumento da variância no SSL, são os dados destoantes de um mesmo conjunto de amostras de solo [4], estes denominados *outliers*.

Os estudos sugerem que estratificar a SSL com base em grupos de solo, textura [3] uso/cobertura do solo [11] e regiões [9], melhore a precisão das previsões do COS ao reduzir a variância do SSL. Porém essas informações quanto a origem da amostra nem sempre estão disponíveis numa SSL, logo uma estratégia é estratificar a SSL com base na variância espectral . Embora a diminuição da variância da SSL, melhore o desempenho da predição dos modelos, estudos relatam que devido a diminuição do conjunto de dados, a acurácia do modelo pode ser reduzida.

Nesse contexto, o objetivo principal deste estudo foi estratificar um banco de dados com 18.716 amostras de solo, por meio da variância espectral com o script de Análise Espectral de Solo de Rotina Não Supervisionada (URSSA) [4], para identificar valores *outliers* de atributos do solo em grupos de amostras espectralmente similares e analisar o efeito da remoção desses *outliers* e do tamanho do conjunto de dados em modelos de predição.

2. MATERIAL E MÉTODOS

2.1 Obtenção dos dados espectrais

Por meio de uma parceria firmada entre a Fundação ABC e o Grupo Geotecnologias em Ciências do Solo (GeoCis), sediado no Departamento de Solos da Escola Superior de Agricultura "Luiz de Queiroz" (ESALQ-USP). Foram enviadas um banco de 18.730 amostras de solo, oriundas do laboratório da fundação ABC (abcLab) que passaram por análises químicas e granulométricas. A origem destas amostras compreende a região do estado do Paraná. E nas dependências do abcLab, um laboratório de análise química de rotina, foram analisados 30 atributos de solo das amostras do banco enviado, dentre eles a Matéria Orgânica (M.O). Todas as amostras de solo foram previamente secas a 45°C, moídas e peneiradas com malha de 2 mm no laboratório abcLab, dessa forma as amostras vieram devidamente preparadas. Posteriormente distribuídas homogeneamente em placas de Petri antes da obtenção dos espectros. Os dados espectrais foram adquiridos utilizando os espectrorradiômetros *Fieldspec* 3, 4 e Pro (*Analytical Spectral Devices*, ASD, *Boulder*, CO), que possui uma faixa espectral do visível ao infravermelho de ondas curtas. (350–2500 nm) e resolução espectral de 1 nm de 350 a 700 nm, 3 nm de 700 a 1400 nm e 10 nm de 1400 a 2500 nm. O intervalo de amostragem da saída de dados é de 1 nm, totalizando 2151 bandas.

2.2 Padronização dos dados espectrais

Fixado o período de calibração do sensor em 20 minutos, a obtenção dos dados espectrais foi sistematizada em lotes de análises, e em cada lote, contou com a análise de duas amostras de areias denominadas padrão, areias oriundas ao longo da costa de *Wylie Bay* (WB) e *Lucky Bay* (LB) no sudoeste da Austrália. Para que dessa forma posteriormente fosse aplicado um protocolo de padronização das curvas espectrais, o qual foi proposto por Ben dor e colaboradores (2015), na estruturação do banco de dados.

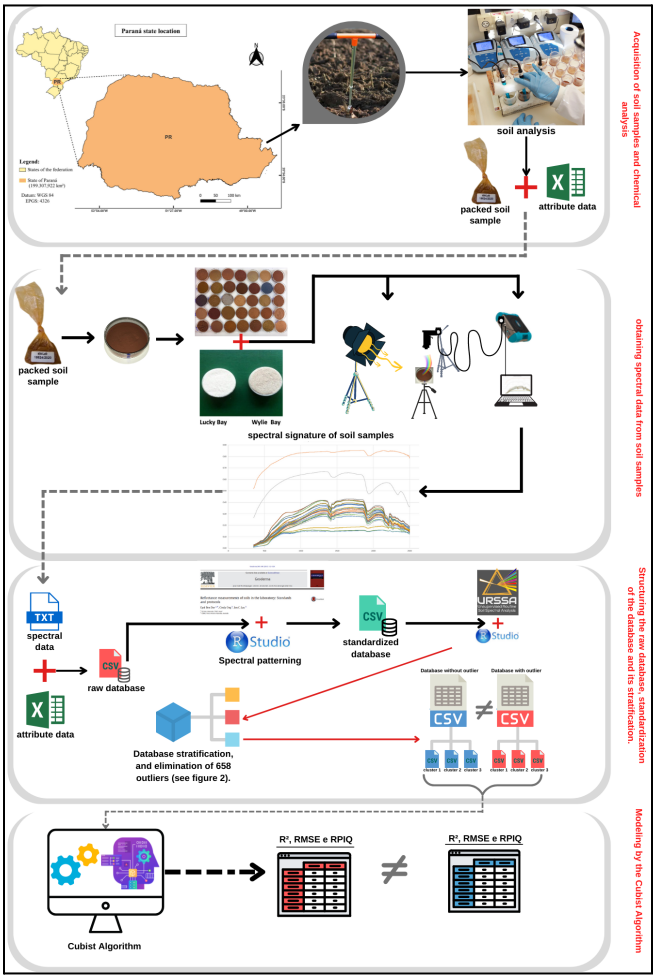


Figura 1. Sequência metodológica.

2.3 Estratificação e identificação dos *outliers*

Posteriormente o script de Análise Espectral de Solo de Rotina Não Supervisionada (URSSA) [4], foi aplicado no banco de dados padronizado, com objetivo de remover *outliers* dos teores de MOS que foram determinados usando métodos analíticos clássicos do banco de dados, e estratificar o banco em *clusters* como parâmetro a variância espectral. O banco de dados foi estratificado em 4° níveis, obtendo ao final 273 *clusters*, o qual nesses subconjuntos de dados foram sinalizadas as amostras com valores de teores de MOS destoantes do *cluster* de nível 4° de estratificação. Com o número de identificação destas amostras, as mesmas foram excluídas do banco de dados e dos *cluster* de 1° nível de estratificação, para posterior modelagem preditiva.

2.4 Modelagem preditiva e análise de variância

Modelos preditivos foram ajustados com os três, banco de dados, utilizando o algoritmo *Cubist* [12] no R. As amostras foram divididas aleatoriamente em um conjunto de calibração (70%) e validação (30%), sendo processadas 50 vezes. Os parâmetros estatísticos avaliados foram o coeficiente de determinação (R^2), raiz quadrática do erro médio (RMSE) e razão da performance para distância interquartil (RPIQ). O teste não paramétrico *Mann-Whitney* foi utilizado para avaliar se houve diferenças estatísticas entre os valores de R^2 dos tratamentos com e sem *outliers*.

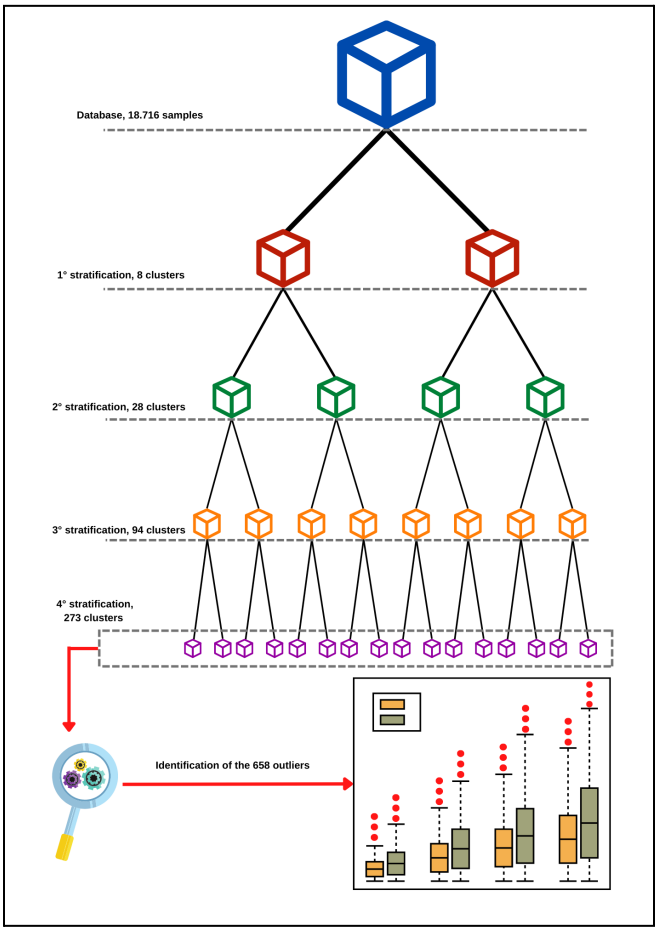


Figura 2. Estratificação do banco de dados e identificação dos *outliers*.

3. RESULTADOS E DISCUSSÃO

A predição da MOS foi superior aos tratamentos com a remoção de *outliers*, para o conjunto do banco de dados inteiro, *cluster* 3 e *cluster* 8, com R^2 iguais a 0.84, 0.87, 0.74 respectivamente e RMSE de 7.43, 9.88, 5.5 $g.kg^{-1}$ respectivamente. No que tange ainda ao tratamento com a remoção de *outliers*, o mesmo foi capaz de diminuir os valores do RMSE para os *clusters* 1, 2, 6 e 7 respectivamente, além de ter melhorado a capacidade do modelo em lidar com a dispersão do conjunto de dados indicados pelo RPIQ, nos *clusters* 1, 2, 5 e 7. A otimização destas métricas foi em função de uma melhor distribuição dos dados dos teores da MOS quantificados em laboratório, como mostra os histogramas (figura 3) depois da remoção dos *outliers* (figura 4). Ficou evidenciado, que os valores de RMSE diminuíram após a estratificação em *clusters* de dados, em comparação a modelagem com o banco de dados inicial. Outro ponto a ressaltar que para os *cluster* com menores números de amostras, houve uma redução no ajuste do modelo, indicados pelo R^2, sendo que no tratamento sem *outlier* este quesito é ainda mais pronunciado, pelo fato do conjunto de dados apresentar menos amostras em relação ao mesmo *cluster* do outro tratamento, sendo esse efeito relatado por Demattê e colaboradores (2016).

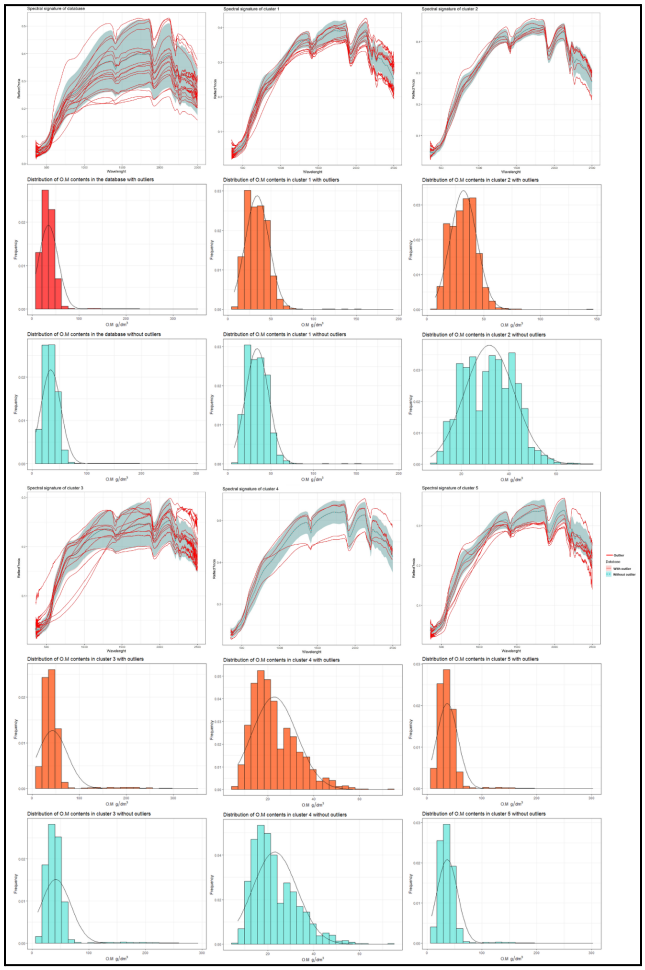


Figura 3. Relação da média e desvio padrão espectral do banco de dados e clusters 1, 2, 3, 4 e 5 , com seus respectivos histogramas do teor de MOS nos tratamentos com e sem *outliers*.

Destacam-se algumas assinaturas espectrais sinalizadas como *outliers*, ilustradas na cor vermelha (figura 3), as quais destoam da média e desvio padrão do *cluster*. Aparentemente, afigura-se um equívoco afirmar que esses espectros são *outliers*, dado que são amostras de solo com conteúdo de MOS e mineralogia distinta das demais do *cluster*, o qual a estratificação neste primeiro nível, não conseguiu alocar tais amostras num *cluster* mais ajustado,

pelo fato, que neste estudo o URSSA foi limitado a estratificar em até 10 *clusters* os conjuntos de dados. Por isso a importância da estratificação em mais níveis, para que ocorra uma agrupação espectral mais ajustada entre os *clusters*, com a finalidade de identificar as reais amostras que o conteúdo da MOS, sejam destoantes das demais do conjunto de dados, e assim excluirmos essas, do conjunto de dados que se pretende realizar a modelagem preditiva. Logo diminuímos a variância da SSL, acarretando num refino dos modelos preditivos, os quais as métricas estão apresentadas na figura 4.

			Mediana					
			Com Outlier			Sem Outlier		
	Nº de amostras	Nº outliers	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ
BD inteiro	18716	658	0.83 b	8.87 b	1.015 b	0.84 a	7.435 a	1.21 a
cluster 1	3450	133	0.7 a	7.96 b	1.13 b	0.675 b	7.395 a	1.22 a
cluster 2	2960	74	0.75 a	5.66 b	1.59 b	0.72 b	5.47 a	1.65 a
cluster 3	4605	166	0.85 b	12.655 b	0.71 b	0.87 a	9.88 a	0.91 a
cluster 4	1515	38	0.69 a	5.54 a	0.9 a	0.61 b	6 b	0.83 b
cluster 5	2873	104	0.81 a	9.365 a	0.885 b	0.78 b	9.57 b	0.835 a
cluster 6	1012	36	0.715 a	6.165 a	1.15 a	0.64 b	6.175 a	1.13 a
cluster 7	710	43	0.665 a	6.455 b	1.24 b	0.63 b	6.16 a	1.3 a
cluster 8	1591	64	0.75 a	5.615 b	1.425 b	0.74 a	5.5 a	1.46 a

Figura 4. Comparação dos parâmetros estatísticos entre os tratamentos.

5. CONCLUSÕES

Este trabalho evidencia a importância da metodologia da predição de atributos de solo por algoritmos preditivos, baseados em técnicas de espectroscopia Vis-NIR-SWIR. Os resultados obtidos com o processo de estratificação com base em similaridade espectral do banco de dados, se mostrou eficaz, no refino dos modelos em lidar com a dispersão do conjunto de dados, e na diminuição do RMSE, dado a exclusão dos *outliers*. E para estudos futuros a fim de melhor ajustar os modelos preditivos, com base na estratificação via similaridade espectral, queremos investigar qual é o menor tamanho populacional de amostras que o *cluster* deve possuir, antes da acurácia do modelo ser reduzida, para que assim possamos aproveitar ao máximo o benefício da diminuição da variância do conjunto de dados.

Por fim, vê se necessário um aprofundamento nas metodologias de predição dos atributos de solo por modelos computacionais, pois os dados obtidos a partir deste estudo, abrem-se novos horizontes para melhorar as técnicas e modelos de análises, a fim de reduzir custos, impactos ambientais e redução do tempo necessário para analisar-se um grande número de amostras de solo.

6. REFERÊNCIAS

[1] Araújo, S.R.; Wetterlind, J.; Demattê, J.A.M.; Stenberg, B. Improving the Prediction Performance of a Large Tropical Vis-NIR Spectroscopic Soil Library from Brazil by Clustering into Smaller Subsets or Use of Data Mining Calibration Techniques. *Eur J Soil Sci*, v *65*, 718–729, 2014.

[2] Demattê, J.A.M.; Bellinaso, H.; Araújo, S.R.; Rizzo, R.; Souza, A.B. Spectral Regionalization of Tropical Soils in the Estimation of Soil Attributes. *Revista Ciência Agronômica*, v *47*, 589–598, 2016.

[3] Wijewardane, N.K.; Ge, Y.; Wills, S.; Loecke, T. Prediction of Soil Carbon in the Conterminous United States: Visible and Near Infrared Reflectance Spectroscopy Analysis of the Rapid Carbon Assessment Project. *Soil Science Society of America Journal*, v *80*, 973–982, 2016.

[4] Poppiel, R.R.; Paiva, A.F. da S.; Demattê, J.A.M. Bridging the Gap between Soil Spectroscopy and Traditional Laboratory: Insights for Routine Implementation. *Geoderma*, v *425*, 2022.

[5] Ross, C.W.; Grunwald, S.; Vogel, J.G.; Markewitz, D.; Jokela, E.J.; Martin, T.A.; Bracho, R.; Bacon, A.R.; Brungard, C.W.; Xiong, X. Accounting for Two-Billion Tons of Stabilized Soil Carbon. *Science of The Total Environment*, v *703*, 2022.

[6] Nocita, M.; Stevens, A.; van Wesemael, B.; Aitkenhead, M.; Bachmann, M.; Barthès, B.; Dor, E. ben; Brown, D.J.; Clairotte, M.; Csorba, A.; et al. Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring. *Advances in Agronomy*, v.*132*, pp.139–159, 2015.

[7] Dotto, A.C.; Dalmolin, R.S.D.; ten Caten, A.; Grunwald, S. A Systematic Study on the Application of Scatter-Corrective and Spectral-Derivative Preprocessing for Multivariate Prediction of Soil Organic Carbon by Vis-NIR Spectra. *Geoderma*, v.*314*, pp.262–274, 2018.

[8] Van Leeuwen, C.C.E.; Mulder, V.L.; Batjes, N.H.; Heuvelink, G.B.M. Statistical Modelling of Measurement Error in Wet Chemistry Soil Data. *Eur J Soil Sci*, v.*73*, 2022.

[9] Demattê, J.A.M.; Dotto, A.C.; Paiva, A.F.S.; Sato, M. v.; Dalmolin, R.S.D.; de Araújo, M. do S.B.; da Silva, E.B.; Nanni, M.R.; ten Caten, A.; Noronha, N.C.; et al. The Brazilian Soil Spectral Library (BSSL): A General View, Application and Challenges. *Geoderma*, v.*354*, 2019.

[10] Viscarra Rossel, R.A.; Behrens, T.; Ben-Dor, E.; Brown, D.J.; Demattê, J.A.M.; Shepherd, K.D.; Shi, Z.; Stenberg, B.; Stevens, A.; Adamchuk, V.; et al. A Global Spectral Library to Characterize the World's Soil. *Earth Sci Rev*, v.*155*, pp.198–230, 2016.

[11] Moura-Bueno, J.M.; Dalmolin, R.S.D.; Horst-Heinen, T.Z.; ten Caten, A.; Vasques, G.M.; Dotto, A.C.; Grunwald, S. When Does Stratification of a Subtropical Soil Spectral Library Improve Predictions of Soil Organic Carbon Content? *Science of The Total Environment*, v.*737*, 2020.

[12] Quinlan, J.R. Learning with continuous classes, in: 5th Australian Joint Conference on Artificial Intelligence. pp. 343–348.https://sci2s.ugr.es/keel/pdf/algorithm/congreso/1992-Quinlan-AI.pdf.

[13] Zomer, R.J.; Bossio, D.A.; Sommer, R.; Verchot, L. v Global Sequestration Potential of Increased Organic Carbon in Cropland Soils., 10.1038/s41598-017-15794-8.