UNIVERSIDADE FEDERAL DO RIO GRANDE
ESCOLA DE QUÍMICA E ALIMENTOS
ENGENHARIA AGROINDUSTRIAL AGROQUÍMICA

ESTUDO DA VIABILIDADE TÉCNICA NA UTILIZAÇÃO DE ESPECTROSCOPIA RAMAN NA CARACTERIZAÇÃO ONLINE DE SOLOS

Lucas do Amaral Freitas

2018

ESTUDO DA VIABILIDADE TÉCNICA NA UTILIZAÇÃO DE ESPECTROSCOPIA RAMAN NA CARACTERIZAÇÃO ONLINE DE SOLOS

Lucas do Amaral Freitas

Projeto de Conclusão de Curso apresentado à Universidade Federal do Rio Grande, como parte dos requisitos necessários à graduação em Engenharia Agroindustrial Agroquímica.

Orientador: Prof. Dr. Cassiano Ranzan
Coorientador: Me. Lucas Ranzan

Santo Antônio da Patrulha, RS

2018

AGRADECIMENTOS

Agradeço as duas mulheres da minha vida, a minha mãe Roselma e minha namorada Valentina, por serem meu porto seguro e me apoiarem durante toda a graduação e terem me suportado quando ninguém mais suportava.

Agradeço ao meu orientador Cassiano Ranzan e ao meu coorientador Lucas Ranzan pelo apoio durante este ano e durante este trabalho de conclusão e por terem paciência e terem me incentivado a investir neste tema.

Agradeço ao Curso de Engenharia Agroindustrial Agroquímica da Universidade Federal do Rio Grande (Campus Santo Antônio da Patrulha), e aos amigos com quem convivi neste espaço ao longo desses anos, todos contribuíram de alguma forma para eu ter chegado até o final.

RESUMO

O crescimento da competição na agricultura e a necessidade de uma produção eficiente acarretaram em um aumento em custos e aplicação indiscriminada em relação aos insumos. Visando aperfeiçoar a aplicação destes insumos surgiu a agricultura de precisão. Para sua utilização é necessário um prévio levantamento das características do solo, sendo sua eficiência limitada pela qualidade desta caracterização. Usualmente o solo é caracterizado por amostragem, quantificadas analiticamente, processo esse oneroso e demorado. Assim sendo, existe a necessidade de se encontrar uma forma rápida, de baixo custo e eficiente de caracterizar o solo. Neste contexto, a espectroscopia aparenta ser uma técnica viável para esta aplicação. Este trabalho propõe o estudo da viabilidade de utilização da espectroscopia Raman para a caracterização online de solos. Para tanto foram utilizados conjuntos de amostras de solo coletadas na região de Santo Antônio da Patrulha, previamente caracterizadas analiticamente de acordo com o normativo pertinente e caracterizadas através de medidas de espectroscopia. Posteriormente, foi realizada a análise em componentes principais e aplicada a metodologia de otimização heurística baseada no algoritmo de Colônia de Formigas (ACO) para filtragem, seleção de quantidades finitas de dados de espectroscopia e criação de modelos para predição das características de interesse nas amostras. Os resultados das análises numéricas indicaram que a espectroscopia Raman é útil se tratando de encontrar correlações entre os conjuntos de dados, mas não possuiu resultados quantitativos significativos para a geração de modelos para caracterização. Porém, a espectroscopia para análise de solos é uma área que pode e deve ser estudada mais a fundo.

Palavras-chaves: agricultura de precisão, espectroscopia Raman, caracterização *online* de solos, viabilidade técnica, otimização colônia de formigas.

ABSTRACT

The growth of competition in agriculture and the need for efficient production led to an increase in costs and indiscriminate application of inputs. Aiming to optimize the application of these inputs came the precision agriculture. For its use it is necessary a previous survey of the characteristics of the soil, where its efficiency limited by the quality of this characterization. Usually the soil is characterized by sampling and analytically, this process costly and time consuming. Therefore, there is a need to find a fast, low cost and efficient way to characterize the soil. In this context, spectroscopy appears to be a viable technique for this application. This work proposes the feasibility study of Raman spectroscopy for the online characterization of soils. Soil samples were collected in the Santo Antônio da Patrulha region, previously characterized analytically according to the relevant regulations and characterized by measurements of Raman spectroscopy. Subsequently, the Principal Component Analysis was applied and the heuristic optimization methodology based on the Ant Colony algorithm (ACO) was applied for filtering, selection of finite amounts of spectroscopy data and creation of models for the characterization of analytical data. Numerical analysis indicated that Raman spectroscopy is useful when trying to find correlations between the data sets, but did not have significant quantitative results for model generation. However, spectroscopy for soil analysis is an area that can and should be further studied.

Keywords: precision agriculture, Raman spectroscopy, online characterization of soils, technical viability, ant colony optimization.

LISTA DE FIGURAS

Figura 1: Fotoluminescência (fluorescência e fosforescência). (a) Absorção de radiação eletromagnética e dissipação de energia por emissão de radiação. (b) variação de energia dos níveis eletrônicos na etapa de absorção e emissão de fótons. (c) espectro correspondente aos comprimentos de onda emitidos associados às diferenças de energia entre os níveis. 16

Figura 2: Na Microscopia Raman a objetiva do microscópio é usada tanto para focalizar o laser sobre a superfície da amostra quanto para coletar a radiação espalhada. 17

Figura 3: Diagrama mostrando a evolução, no tempo, do experimento de Goss et al., (1989) sobre o estudo do processo de busca de comida pelas formigas da espécia L. humilis, onde o caminho entre o ninho e a fonte de comida é otimizado através do trabalho conjunto da colônia. 21

Figura 4: Representação esquemática do algoritmo ACO para seleção de grupos de elementos espectrais. 22

Figura 5: Representação da metodologia proposta na forma de diagrama. 28

Figura 6: Representação gráfica dos dados de espectroscopia Raman médio obtidos para o conjunto amostral em estudo. 30

Figura 7: Análise de PCA do conjunto bruto de dados analisados. 32

Figura 8: Análise de PCA da média dos espectros para o conjunto de dados analisados. 33

LISTA DE TABELAS

Tabela 1: Caracterização quantitativa laboratorial do conjunto amostral, fornecido pelo Labsolos FURG/SAP, sendo 1 cmolc igual a 10 mg de hidrogênio. 29

Tabela 2: Melhores resultados obtidos em cada parâmetro na prospecção primária da capacidade preditiva dos modelos. 35

Tabela 3: Melhor modelo ajustado para caracterização do parâmetro de pH, obtido com 200 iterações e 400 formigas. 36

SUMÁRIO

1. INTRODUÇÃO 10
2. OBJETIVOS 12
2.1. Objetivo Geral 12
2.2. Objetivos Específicos 12
3. REFERENCIAL TEÓRICO 13
3.1. Seleção de amostras e caracterização analítica 13
3.1.1. Acidez potencial e ativa do solo (H+Al) 13
3.1.2. Fósforo e Potássio 13
3.1.3. Matéria orgânica 14
3.1.4. Alumínio, Cálcio e Magnésio 14
3.1.5. Capacidade de troca de cátions 14
3.2. Caracterização através de espectroscopia Raman 15
3.3. Tratamento de dados de espectroscopia 17
3.3.1. Análise Qualitativa – Análise de Componentes Principais (PCA) 17
3.3.2. Normalização dos dados 19
3.4. Análise multivariável 19
3.5. Seleção e ajuste de modelos quimiométricos lineares para predição das variáveis analíticas 20
3.6. Validação de modelos 22
4. METODOLOGIA 24
4.1. Seleção de amostras de solo e caracterização analítica 24
4.2. Caracterização através de espectroscopia Raman 24
4.3. Tratamento de dados de espectroscopia 25
4.3.1. Normalização dos dados 25
4.3.2. Analise Qualitativa – Análise de Componentes Principais 26
4.4. Análise multivariável 26
4.4.1. Seleção e ajuste de modelos quimiométricos lineares para predição das variáveis analíticas 27
4.5. Validação de modelos 27
4.6. Visão geral da metodologia 28
5. RESULTADOS E DISCUSSÃO 29
5.1. Amostras e Caracterização Laboratorial 29
5.2. Dados de Espectroscopia Raman 30
5.3. Análise de Componentes Principais (PCA) 31
5.4. Análise Multivariável (ACO) 33
6. CONSIDERAÇÕES FINAIS 38

CRONOGRAMA DE ATIVIDADES 40
REFERÊNCIAS BIBLIOGRÁFICAS 41

1. INTRODUÇÃO

Atualmente a agricultura tem se tornado uma atividade cada vez mais competitiva no aspecto econômico, o que fomenta a produção eficiente, onde se busca o ótimo de produtividade, significando, produzir mais com alta qualidade e em extensões de área limitadas. Visando o aumento da produção, a utilização de insumos, como fitossanitários ou fertilizantes, tornou-se mandatório, fato este que impacta diretamente nos custos de produção e traz uma nova problemática, que é a definição de quais, onde, quando e em que quantidade de insumos utilizar (MOLIN, 2017).

Usualmente, a aplicação dos insumos é feita de forma indiscriminada, baseada em valores médios de referência para características regionais. Esta estratégia, por usar dados médios, acarreta em resultados não satisfatórios. Estes resultados são oriundos da utilização elevada em determinadas regiões e em utilização aquém do necessário em outras (seja no espaço, no tempo ou na espécie). Esta estratégia, de forma geral, leva a uma produção não ótima e desuniforme, com desperdício e não otimalidade da produtividade ou lucratividade.

A agricultura de precisão surgiu com o objetivo de solucionar o problema de otimalidade de produção em relação a utilização de recursos. Esta estratégia parte de uma prévia caracterização discretizada do solo e constrói um plano de utilização e aplicação de recursos otimizado, projetado de forma individualizada, visando assim, minimizar desperdícios e maximizar a efetividade dos insumos aplicados (RESENDE et al., 1999).

A agricultura de precisão necessita de uma classe de informações à priori, que viabiliza sua aplicação. Estas informações são a caracterização quantitativa das áreas de produção, ou seja, as características físico-químicas do solo onde se deseja produzir. Conhecendo as características do solo e as necessidades do cultivar que se deseja produzir, a agricultura de precisão projeta a melhor estratégia de utilização de recursos para o aperfeiçoamento da produção. Portanto, a efetividade do projeto de agricultura de precisão está diretamente relacionada à qualidade (precisão, exatidão, quantidade, frequência, etc.) da caracterização do solo (PINHEIRO, 2016).

Atualmente, o processo de caracterização do solo em áreas produtivas é feito através de coleta de amostras de solo e análise laboratorial das mesmas. As amostras são coletadas através de procedimentos padronizados e em posições pré-determinadas, de forma a garantir uma amostragem representativa da área, mas que utilize a menor quantidade possível de amostras. Sendo assim, o processo de caracterização do solo, para aplicação das técnicas de

agricultura de precisão, é demasiadamente trabalhoso, dispendioso e demorado (BALASTREIRE, 2000).

É de interesse do ramo agroindustrial o desenvolvimento de uma técnica confiável, rápida e barata para a caracterização do solo. Uma técnica com estas características permitiria um acompanhamento mais constante das características físico-químicas do solo, além de permitir discretizações menores das áreas produtivas, melhorando a efetividade da agricultura de precisão.

Pensando em técnicas de caracterização de amostras, as técnicas baseadas em emissão e recepção de luz mostram-se com mais aplicabilidade para este fim. Espectroscopia é um conjunto de técnicas que consistem na medida da quantidade de radiação absorvida ou emitida por espécies de interesse. A espectroscopia é amplamente empregada na determinação qualitativa e quantitativa de compostos orgânicos e inorgânicos. Além disso, é um método de análise barato, rápido, não destrutivo, não intrusivo e com tempo de resposta pequeno. Todas estas características indicam sua viabilidade de utilização em sistemas embarcados para a caracterização de solos (FARIA, 2011).

Apesar da inúmera quantidade de vantagens associadas à utilização de métodos de espectroscopia para a caracterização de processos, mais especificamente de amostras de solo, é necessário, primordialmente, determinar se dentre as técnicas de espectroscopia existentes, alguma delas é capaz de caracterizar amostras de solo frente às variáveis de interesse da agricultura de precisão. É necessário, também, determinar quantos e quais elementos espectrais devem ser utilizados nesse processo, em que estrutura de modelo devem ser combinados e como ajustar e testar estes modelos.

Sendo assim, o presente trabalho busca estudar a viabilidade técnica da utilização da espectroscopia Raman para caracterização online de solos para tornar o mapeamento de solos mais rápido.

2. OBJETIVOS

2.1. Objetivo Geral

O presente trabalho tem por objetivo estudar a viabilidade de utilização de Espectroscopia Raman para a caracterização rápida de amostras de solo.

2.2. Objetivos Específicos

- Comparar qualitativamente, através da metodologia quimiométrica de PCA, os dados espectrais com os dados analíticos das amostras previamente analisadas;
- Definir uma estratégia para segmentação do conjunto amostral em conjunto de amostras para calibração e conjunto de amostras de teste de modelos preditivos;
- Aplicar a metodologia de otimização estocástica de Colônia de Formigas (ACO) para selecionar componentes espectrais individuais e combina-los, através de modelos lineares nos parâmetros, para a predição das propriedades analíticas disponíveis;
- Avaliar e comparar a capacidade preditiva dos modelos quimiométricos obtidos pelos critérios da Raiz Quadrática Erro Padrão Médio (RMSE), Erro Médio Percentual Absoluto (MAPE) e do Coeficiente de Determinação (R^2).

3. REFERENCIAL TEÓRICO

3.1. Seleção de amostras e caracterização analítica

O conjunto de amostras selecionado para este trabalho é composto por alíquotas de solo coletadas de 10 propriedades locais de Santo Antônio da Patrulha. As referidas amostras foram cedidas pelo Laboratório de Análise de Solos (Labsolos), localizado na unidade Bom Princípio, sala 9, do campus FURG/SAP, sob supervisão do professor Carlos Peixoto, que por sua vez, recebeu-as através do Sindicato dos Trabalhadores Rurais de Santo Antônio da Patrulha.

A caracterização analítica das amostras foi realizada pelo Labsolos seguindo o manual elaborado por Tedesco (1995). A seguir são descritas as principais características de qualidade de solo avaliadas.

3.1.1. Acidez potencial e ativa do solo (H+Al)

A acidez do solo pode ser dividida em acidez ativa e acidez potencial, e esta, por sua vez, em acidez trocável e acidez não trocável. Denomina-se acidez ativa a parte do hidrogênio que está dissociada, ou seja, na solução do solo, na forma de H^+ e é expressa em valores de pH. Nos solos (à semelhança aos ácidos fracos) a maior parte do hidrogênio não está dissociada. A acidez trocável refere-se aos íons H^+ e Al^{3+} que estão retidos na superfície dos colóides por forças eletrostáticas.

A quantidade de hidrogênio trocável, em condições naturais, parece ser pequena. A acidez não trocável é representada pelo hidrogênio de ligação covalente, associado aos colóides com carga negativa variável e aos compostos de alumínio. A acidez potencial corresponde à soma da acidez trocável e da acidez não trocável do solo (LOPES, 1991).

A maior preocupação do agricultor deve ser em corrigir a maior parte da acidez potencial, que é a mais prejudicial ao crescimento da maioria das plantas.

3.1.2. Fósforo e Potássio

O fósforo (P) é considerado um nutriente de baixa mobilidade no solo, comportamento atribuído à sua “fixação” pelos minerais da argila, e esse elemento tem presença relevante nos solos tropicais que apresentam elevados teores de óxidos de ferro e de alumínio – com os quais o P tem grande afinidade. Normalmente, entre 20% e 30% do fósforo aplicado como fertilizante é aproveitado pelas culturas anuais em solos tropicais, sendo necessária a

aplicação de quantidades que, em geral, superam em muito as extrações dessas culturas (PEREIRA, 2009).

O potássio (K) do solo é, usualmente, distinguido nas seguintes formas: K na solução do solo (Ks); K trocável (Kt), fracamente retido na capacidade de troca de cátions (CTC) do solo; K não trocável (K_{nt}), retido na estrutura de minerais (K estrutural), mais o retido nas entrecamadas de argilominerais expansivos (K "fixado"); K total (KT), extraído com HF. Existe certo equilíbrio entre essas formas de potássio no solo. As plantas absorvem K da solução, o qual é tamponado pelas formas trocáveis, que são repostas pelas formas não trocáveis e estruturais.

As formas não trocáveis e estruturais de K são usualmente consideradas reservas de médio e longo prazo para as plantas. Sob o ponto de vista da nutrição de planta, o equilíbrio mais importante se dá entre K trocável e K solução (cujo somatório, para fins práticos, é considerado K "disponível"), que são as fontes imediatas de K para as plantas (KÄMPF; MARQUES; CURI, 2012).

3.1.3. Matéria orgânica

Formada por organismos, resíduos de vegetais e de animais em decomposição. Além do teor de argila, a matéria orgânica do solo é a principal reguladora da CTC, proporcionando uma maior retenção de cátions (K, Ca, Mg), evitando que sejam lixiviados e suprindo as plantas de nutrientes (SALLES, 2015).

3.1.4. Alumínio, Cálcio e Magnésio

O cálcio e o magnésio são os principais componentes da CTC na maioria dos solos (TEDESCO, 1995). Enquanto a Saturação de Alumínio é a quantidade de cargas negativas do solo que está ocupada pelo Al^3 trocável, próximo ao pH natural do solo, expressando a toxidez do alumínio no solo. Quanto mais ácido for o solo, maior poderá ser o teor de alumínio trocável, maior a percentagem de saturação por Al, menores os teores de Ca, Mg, K e, consequentemente, menor a soma de bases trocáveis (SALLES, 2015).

3.1.5. Capacidade de troca de cátions

É definida como sendo a soma total dos cátions que o solo pode reter na superfície coloidal prontamente disponível à assimilação pelas plantas. Solos com menor valor de CTC

requerem menor quantidade de calcário para aumentar o pH, o nitrogênio e o potássio lixiviam mais e tem menor capacidade de retenção de umidade (SALLES, 2015).

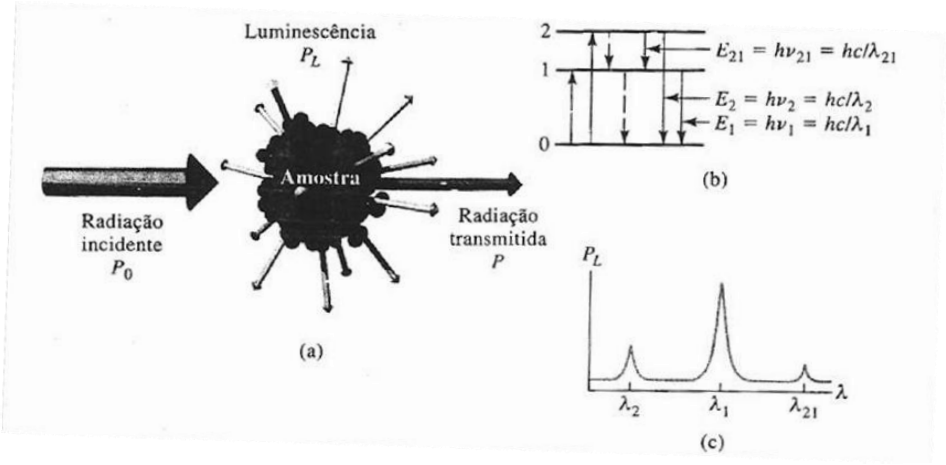
3.2. Caracterização através de espectroscopia Raman

Medidas baseadas na luz e outras formas de radiação eletromagnética são amplamente empregadas em caracterização de soluções e processos. As interações da radiação com a matéria são o objeto de estudo da espectroscopia. Métodos espectroscópicos baseiam sua análise na medida da quantidade de radiação produzida ou absorvida por moléculas ou espécies atômicas de interesse. Estes métodos podem ser classificados de acordo com a região do espectro eletromagnético envolvido na medida (SETTLE, 1997).

Em análises espectroscópicas, a amostra é estimulada aplicando-se energia na forma de calor, energia elétrica, luz, partículas ou por uma reação química. Previamente ao estímulo, o analito se encontra predominantemente em seu estado fundamental. O estímulo então faz com que algumas espécies do analito sofram uma transição para um estado excitado de energia. Informações são obtidas através da análise da radiação emitida quando se retorna ao estado fundamental ou a quantidade de radiação eletromagnética absorvida decorrente da excitação.

A Figura 1 ilustra o fenômeno responsável pela espectroscopia de fotoluminescência, onde a emissão de fótons é medida após a absorção da energia incidente. As formas mais importantes de fotoluminescência para os propósitos analíticos são as espectroscopias de fluorescência e fosforescência.

Figura 1: Fotoluminescência (fluorescência e fosforescência). (a) Absorção de radiação eletromagnética e dissipação de energia por emissão de radiação. (b) variação de energia dos níveis eletrônicos na etapa de absorção e emissão de fótons. (c) espectro correspondente aos comprimentos de onda emitidos associados às diferenças de energia entre os níveis.



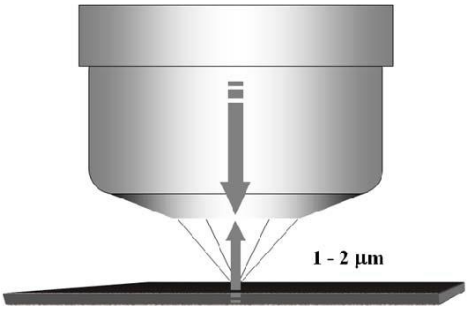
Fonte: Adaptado de Skoog, et al. (2007).

A espectroscopia Raman trata-se de uma técnica que usa uma fonte monocromática de luz a qual, ao atingir um objeto, é espalhada por ele, gerando luz de mesma energia ou de energia diferente da incidente.

Um feixe de radiação laser de baixa potência é usado para iluminar pequenas áreas do objeto de interesse e ao incidir sobre a área definida, é espalhado em todas as direções, sendo que uma pequena parcela dessa radiação é espalhada inelasticamente, com frequência (ou comprimento de onda) diferente da incidente ($E=h\nu$ ou $E=h.c.\lambda^{-1}$). Esse fenômeno foi observado experimentalmente em 1928 por Chandrasekhara Venkata Raman, na Índia e, por esse motivo, foi chamado de efeito Raman.

Caso seja utilizado um microscópio óptico (Figura 2) convencional, no qual a objetiva tanto serve para focalizar o feixe incidente na amostra quanto para coletar a radiação que é espalhada por ela, tem-se a Microscopia Raman, a qual permite o estudo de áreas de até 1 μm (10^{-6} m) de diâmetro (FARIA, 2011).

Figura 2: Na Microscopia Raman a objetiva do microscópio é usada tanto para focalizar o laser sobre a superfície da amostra quanto para coletar a radiação espalhada.



Fonte: Faria (2011).

A análise por Espectroscopia Raman é feita sem necessidade de preparações ou manipulações de qualquer natureza, como pulverização por exemplo. Quando a amostra é muito grande, ela pode ser examinada através de um sistema especial de lentes ou através de fibras ópticas. Atualmente são comercializados tanto microscópios Raman dedicados quanto equipamentos convencionais que podem ter o microscópio como acessório e em ambos os casos a utilização de fibras ópticas é um recurso essencial quando o objeto analisado é muito grande para caber no compartimento de amostras ou quando seu transporte até o laboratório não é possível (CLARK; DINES, 1986).

Para se realizar medidas de espectroscopia Raman são necessários, basicamente, três componentes: uma fonte de luz excitadora, um sistema dispersivo capaz de decompor a radiação espalhada pela amostra e um fotodetector (RODRIGUES; GALZERANI, 2012).

Por ser um processo rápido, não destrutivo, rico em informação, não intrusivo e com alta sensibilidade, a espectroscopia em geral, incluindo a Raman, pode ser vista como uma alternativa para a caracterização de solos.

3.3. Tratamento de dados de espectroscopia

3.3.1. Análise Qualitativa – Análise de Componentes Principais (PCA)

PCA é uma ferramenta que busca encontrar similaridades ou diferenças em amostras de um determinado conjunto de dados. Este método define novas variáveis, obtidas através de combinações lineares das originais. O resultado é um conjunto de novos vetores, ortogonais entre si, chamados de componentes principais (PC – *Principal Component*), resultante de um agrupamento de variáveis altamente correlacionadas. O PCA, então, reduz a dimensionalidade

do conjunto de dados original, fazendo com que as informações mais relevantes fiquem concentradas nos primeiros componentes principais (REDDY, 2011).

O algoritmo utilizado para o cálculo dos PCs é a Decomposição em Valores Singulares (SVD – *Singular Value Decomposition*). O SVD decompõe uma matriz de dados ***X***, de dimensões *n* x *p* (sendo *n* linhas e *p* colunas), centralizada na média, em três partes de acordo com a Equação 1. Os resultados das duas primeiras matrizes (*U* e *D*) são chamados de *scores* na terminologia da PCA. A variável *a* indica o número de componentes principais, sendo equivalente ao menor número entre linhas e colunas da matriz *X* ($a_{max} = min(n, p)$**)**.

$$X = (UD)V^T = TP^T \tag{1}$$

A interpretação das matrizes *T*, *P*, *U*, *D* e *V* é direta. Os *loadings*, colunas da matriz *P* (ou equivalentemente, os vetores singulares à direita, colunas na matriz *V*) fornecem os pesos das variáveis originais nos PCs. As colunas em *U* fornecem as mesmas coordenadas em uma forma normalizada (variância unitária), enquanto as colunas em *T* possuem variância correspondente as variâncias equivalentes a cada PC. Estas variâncias, λ_i, são proporcionais aos quadrados dos elementos diagonais da matriz *D*, como mostra a Equação 2, onde d_i representa o elemento *i*, diagonal da matriz *D*.

$$\lambda_i = \frac{d_i^2}{(n-1)} \tag{2}$$

A variância explicada (*FV***)** de cada PC pode ser obtida de acordo com a Equação 3 (CAMACHO; PICÓ; FERRER, 2010).

$$FV(i) = \frac{\lambda_i}{\sum_{j=1}^{a} \lambda_j} \tag{3}$$

Estes novos dados correspondem a novas amostras, medidas ou instrumentos. O objetivo é que a representação em espaços dimensionais menores permita uma avaliação dos dados de forma mais ampla e simples para a identificação de padrões, possível através da obtenção dos *Scores* do novo conjunto de dados. Partindo de uma nova matriz de dados ***X***, a projeção desta no espaço definido pelos *Loadings* (*P*), pode ser obtida de acordo com a Equação 4.

$$XP = TP^TP = T \tag{4}$$

Um detalhe importante é que o escalonamento da nova matriz de dados deve ser feita com as médias e desvios padrões utilizados na análise de PCA inicial, uma vez que ambos os conjuntos de dados devem possuir o mesmo ponto de origem.

3.3.2. Normalização dos dados

A normalização tem como função fazer pré-tratamento associado a matrizes de dados espectrais, ajustando-os para seleção de variáveis e ajustes matemáticos (RANZAN, 2014a). Este método escalona os dados espectrais de forma que após a normalização a média do espectro seja igual a zero e desvio padrão unitário. A normalização é realizada de acordo com a Equação 5:

$$Zi = \frac{(yi - \bar{y})}{s} \tag{5}$$

Onde y_i é a variável a ser normalizada, $\bar{y}$ valor médio e s o desvio padrão.

3.4. Análise multivariável

Partindo de que o objetivo é determinar o modelo linear dos parâmetros é capaz de converter a matriz de dados mensurados (***X***) na matriz de variáveis preditas (***Y***) é necessário determinar os valores adequados dos parâmetros do modelo (***b***). Então, as equações observacionais podem ser expressas, para o caso generalizado multivariável, na forma da Equação 6 (NASCIMENTO et al., 2010).

$$y = b_0 + x_1 b_1 + x_2 b_2 + \cdots + x_k b_k + e \tag{6}$$

O modelo presente na equação 6 pode ser expandido para sua forma matricial de acordo com a Equação 7, ao se considerar *k* variáveis explicativas (independentes) e *n* observações (amostragens).

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \ddots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix} * \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_k \end{pmatrix} \tag{7}$$

A Equação 7 pode ser escrita de forma compacta de acordo com a Equação 8.

$$y = Xb + e \tag{8}$$

A resolução do problema se encontra na determinação do vetor de parâmetros do modelo (***b***), alcançada através da minimização da soma do quadrado dos resíduos (**min(** $\sum_{i=1}^{n} e_i^2$ **)=(min(**$e^T e$**)**). Após derivar a função objetivo do problema de otimização em relação as variáveis de decisão e igualando esta derivada a zero, é possível chegar à resolução analítica do problema, onde o vetor *b* pode ser calculado de acordo com a Equação 12, e a matriz de variância-covariância dos parâmetros é calculada de acordo com a Equação 13. O termo σ^2 é chamado de variância residual, sendo estimado pela Equação 14.

$$Fobj = e^T e \tag{9}$$

$$e = y - Xb \tag{10}$$

$$\frac{dFobj}{db} = 0 \tag{11}$$

$$b = (X^T X)^{-1} X^T y \tag{12}$$

A variância do vetor *b* pode ser estimada por:

$$Var(b) = (X^T X)^{-1} \sigma^2 \tag{13}$$

Onde σ^2 é definido por:

$$\sigma^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{14}$$

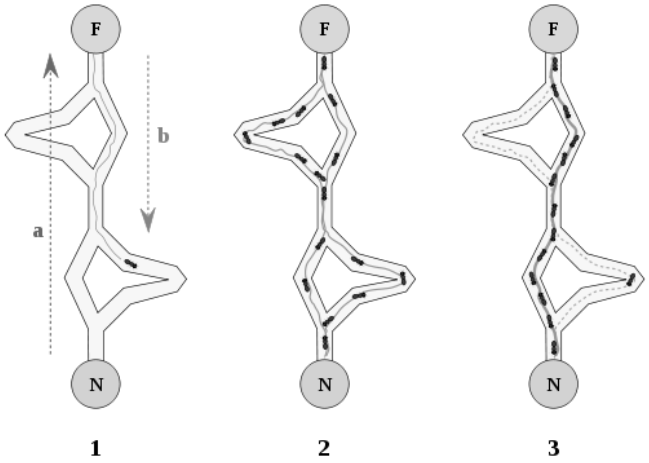
O desvio padrão individual dos coeficientes, fornecido pela raiz quadrada dos elementos diagonais da matriz de variância-covariância (Equação 13), pode ser utilizado para teste estatístico.

3.5. Seleção e ajuste de modelos quimiométricos lineares para predição das variáveis analíticas

Recentemente, a técnica de Otimização Colônia de Formigas (ACO – *Ant Colony Optimization*), que é um método de otimização global não determinístico, tem sido aplicada para resolução de problemas envolvendo seleção de variáveis. É um algoritmo de otimização baseado no comportamento coletivo de formigas quando saem em busca de fontes de alimento. (ALLEGRINI; OLIVIERI, 2011).

A Figura 3 representa um experimento onde as formigas “descobrem” de forma coletiva qual o menor caminho entre o formigueiro e a fonte de alimentos. Foi constatado que isto ocorre devido ao fato das formigas, quando em movimento, secretarem no solo certa quantidade de feromônio, marcando seu trajeto.

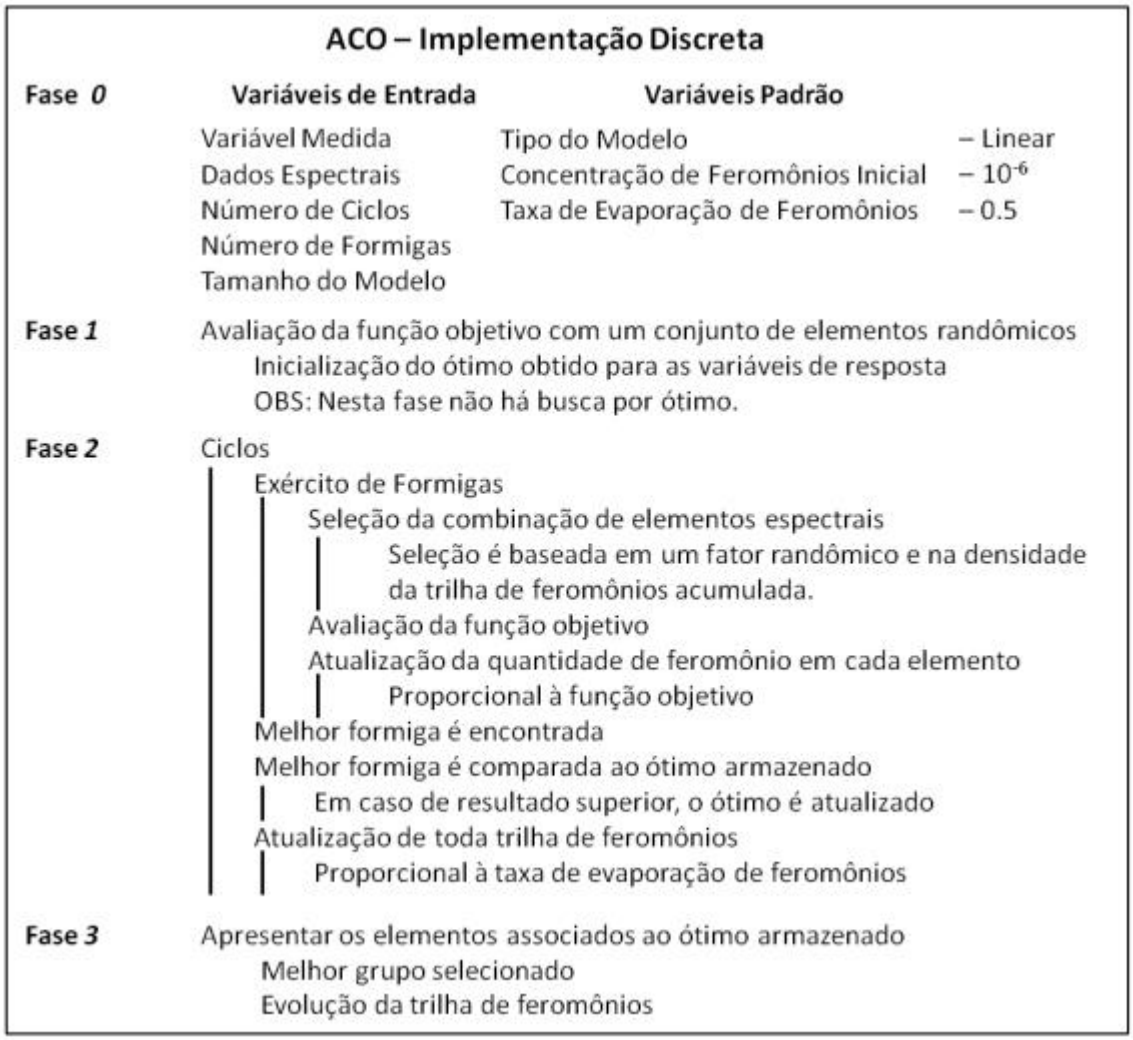
Figura 3: Diagrama mostrando a evolução, no tempo, do experimento de Goss et al., (1989) sobre o estudo do processo de busca de comida pelas formigas da espécia L. humilis, onde o caminho entre o ninho e a fonte de comida é otimizado através do trabalho conjunto da colônia.



Fonte: Adaptado de Goss et al., (1989)

O cerne do algoritmo é exemplificado pela Figura 4, onde o funcionamento do algoritmo é dividido em quatro etapas. A fase 0 se caracteriza pela inicialização de variáveis necessárias para a resolução do problema de otimização. Na fase 1 é inicializado o vetor de soluções, através da resolução da função objetivo com a seleção aleatória de um conjunto de elementos espectrais. A fase 2 é o núcleo do algoritmo de otimização, onde é determinada a melhor combinação de elementos espectrais para predizer a variável de estado observada. Por fim, ocorre a fase 3, onde é apresentado o resultado da melhor solução obtida durante o processo de otimização. Esta seleção corresponde à melhor solução testada pelo exército de formigas (RANZAN, 2014).

Figura 4: Representação esquemática do algoritmo ACO para seleção de grupos de elementos espectrais.



Fonte: Ranzan (2014).

3.6. Validação de modelos

Validação de modelos trata da avaliação da qualidade preditiva de modelos. Um modelo capaz de predições acuradas é considerado um modelo consistente. Validação, entretanto, não é uma tarefa simples. Sistemas químicos ou bioquímicos normalmente apresentam variações significativas, não relacionadas aos parâmetros dos modelos.

Uma vez que o grupo de amostras disponível pode ser considerado representativo, deve ser adotada uma forma de estimar a qualidade das predições, simulando através dos dados coletados, dados que serão preditos.

Para comparar e avaliar os modelos, são necessários métricas para quantificação da qualidade preditiva dos modelos, dentre os quais a raiz quadrada do erro médio de predição

(RMSE – *Root-Mean-Square Error*), erro absoluto percentual médio (MAPE – Mean Absolute Percentage Error), que seria uma estimativa de tendência, na qual geralmente expressa a acurácia do erro em percentagem, definido pela Equação 15, e o coeficiente de determinação R^2 são os mais utilizados (DE MYTTENAERE et al., 2016).

$$MAPE = \frac{100}{n}\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (15)$$

RMSE é calculado de acordo com a Equação 16, onde $\hat{y}$ representa a variável predita pelo modelo, y a variável medida e n o número de amostras contidas conjunto amostral. Este critério analisa o ajuste do modelo ao conjunto de dados, avaliando a reprodutibilidade dos dados.

$$RMSE = \sqrt{\sum_i^n (\hat{y}_i - y_i)^2 / n} \quad (16)$$

O segundo critério para avaliação de modelos é o R^2, que indica o quanto modelos estatísticos se ajustam a dados reais. Este indicador fornece a informação do quanto os dados observados são replicados pelo modelo, quantificando a proporção da variação da resposta que é explicada pelas regressões em um modelo.

O cálculo do coeficiente de determinação é feito de acordo com a Equação 17, onde $\hat{y}$ é a média do vetor de dados medidos (y) e y_{pred} o valor predito pelo modelo (EVERITT, 2002).

$$R^2 = \frac{1 - \sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \bar{y}_i)^2} \quad (17)$$

4. METODOLOGIA

O presente trabalho baseia sua metodologia acerca da aplicação da técnica de espectroscopia Raman na caracterização de amostras de solo. O objetivo desta seção é apresentar as etapas necessárias para o desenvolvimento deste.

A seguir são descritas as etapas vislumbradas para o presente trabalho.

4.1. Seleção de amostras de solo e caracterização analítica

O conjunto de amostras selecionado para este trabalho é composto por alíquotas de solo coletadas de 10 propriedades locais de Santo Antônio da Patrulha. As referidas amostras foram cedidas pelo Laboratório de Análise de Solos (Labsolos), localizado na unidade Bom Princípio, sala 9, do campus FURG/SAP, sob supervisão do professor Carlos Peixoto, que por sua vez, recebeu-as através do Sindicato dos Trabalhadores Rurais de Santo Antônio da Patrulha.

As amostras entregues ao Labsolos foram submetidas ao processo de caracterização físico química, seguindo a metodologia descrita por Tedesco (1995). Cada uma das amostras foi caracterizada segundo seu conteúdo para os seguintes critérios físico-químicos listados abaixo:

- pH em água;
- Argila;
- Fósforo;
- Potássio;
- Matéria orgânica;
- Alumínio;
- Cálcio;
- Magnésio;
- Acidez potencial do solo (H+Al);
- Capacidade de troca de cátions.

De posse dos laudos de caracterização das 10 amostras, foi feita a compilação dos resultados da etapa de caracterização analítica. Estes dados foram considerados como padrões *off-line* e utilizados nas etapas posteriores de calibração e validação de modelos quimiométricos.

4.2. Caracterização através de espectroscopia Raman

As amostras contidas no conjunto amostral, e previamente caracterizadas analiticamente, foram submetidas ao processo de caracterização utilizando a técnica de Espectroscopia Raman.

Cada uma das 10 amostras de solo, foi submetida a coleta de espectro Raman, com o equipamento iHR550 Raman (Horiba Jobin Yvon S.A.S., França) acoplado a um detector de dispositivo de carga acoplada (CCD) e uma fonte de laser de 532 nm.

A potência de saída do laser foi ajustada para 50 mW, e o feixe de laser foi atenuado com um filtro de densidade neutra com 50% de transmitância. Depois disso, o feixe foi refletido por espelhos para um filtro de borda dielétrica. O feixe foi direcionado para o microscópio equipado com uma objetiva de 50X.

O feixe Raman transmitido foi refletido por espelhos de prisma para uma grade de difração de 1800 ranhuras/mm e finalmente para a câmera CCD. Os dados espectrais foram registrados a uma resolução espectral de 0,80 cm^{-1} entre 200 e 4000 cm^{-1} com parâmetros de medição constantes.

Três acumulações de um tempo de contagem de 10 segundos foram utilizadas. O pacote de software LabSpec 5 (Horiba Jobin Yvon S.A.S., França) foi usado para aquisição de dados espectrais. Cada amostra foi mensurada em quatro pontos distintos, desta forma, o conjunto de dados espectrais é composto por 40 espectros.

Os dados de espectroscopia foram condensados em uma única tabela, onde nas linhas estão referenciadas as amostras e nas colunas o resultado de intensidade de emissão associada a cada número de onda avaliado pelo equipamento.

4.3. Tratamento de dados de espectroscopia

A partir do tratamento dos dados de espectroscopia até a etapa 4.5, foi utilizado o algoritmo implementado em Python pelo grupo GIMSCOP/UFRGS, o qual foi adaptado para o trabalho em questão. O mesmo foi executado em um computador com sistema operacional Windows 7 Professional, com um processador Intel® Core™ i3-5005U CPU @ 2.00GHz, com 4,00 GB de memória RAM e operando em 64 bits e a versão do Python utilizada foi a 3.6.6 64 bits, Qt 5.9.3, PyQt5 5.9.2.

4.3.1. Normalização dos dados

Antes da aplicação da técnica de PCA foi necessário tratar os dados, basicamente escalonando os mesmos para a mesma posição espacial e variabilidade. Métodos de pré-processamento especializados são amplamente aplicados em dados de espectroscopia. Um método de normalização usualmente aplicado, e que foi utilizado neste trabalho, é o SNV (*Standard Normal Variate*).

Este método essencialmente faz o escalonamento nas amostras. Isto é, cada espectro terá, após o escalonamento, média 0 e desvio padrão 1. Isto fornece dados livres de *offsets* e fatores de multiplicação. Entretanto, a suposição de que todo espectro deva ter a mesma média e desvio padrão não é sempre válida e pode vir a mascarar dados contidos nos espectros, entretanto, neste trabalho estes fatores foram desconsiderados (WEHRENS, 2011).

4.3.2. Analise Qualitativa – Análise de Componentes Principais

Dados de espectroscopia contém uma quantidade significativa de informações, entretanto, altamente correlacionadas. Sendo assim, antes de iniciar o processo de modelagem utilizando estes dados foi necessário proceder seu tratamento para determinar as melhores condições de análise (WEHRENS, 2011).

Qualitativamente foi utilizado um método de análise não supervisionado que é a Análise de Componentes Principais (PCA – *Principal Componente Analysis*) com o intuito de caracterizar os dados de espectroscopia. Nesta técnica existe a separação de classes baseada na similaridade estatística dos dados espectrais, sem a necessidade de informações sobre a natureza das amostras (dados *off-line*). Dentre os objetivos desta etapa está a identificação de grupamentos naturais (formação de *clusters*) e identificação de tendências (BRERETON, 2009).

4.4. Análise multivariável

De posse do conjunto de dados *off-line* e dos dados de espectroscopia normalizados, foi realizada a análise multivariável dos mesmos, com o intuito de avaliar a possibilidade de caracterização *on-line* de amostras de solo através de modelos multivariáveis e dados de espectroscopia Raman.

A análise de dados multivariáveis consiste no exame de um elevado número de variáveis, de forma a determinar toda a variabilidade da matriz de espectros. O conceito de análise de variação consiste na filtragem e seleção da informação útil contida nessa matriz, descartando ruídos e demais informações que não são úteis ou relativas ao objeto de estudo almejado (BOSQUE-SENDRA et al., 2012).

Para tanto, são gerados modelos multivariáveis, lineares nos parâmetros, e ajustados através do método conhecido como Regularização Tikhonov, ou Regressão Ridge. Esta regularização é uma técnica para analisar dados de regressão múltipla que sofrem de multicolinearidade. Quando ocorre a multicolinearidade, as estimativas de mínimos quadrados

são imparciais, mas suas variações são grandes, de modo que podem estar distantes do valor verdadeiro. A motivação inicial da regressão de Ridge foi incluir uma perturbação na diagonal principal de X^TX de forma que essa possa ser invertida mesmo no caso de ser singular. Ao adicionar um grau de viés às estimativas de regressão, reduz os erros padrão. Espera-se que o resultado seja fornecer estimativas mais confiáveis (FEARN, 2013).

Pela Equação 18, pode-se encontrar um valor de sintonia λ com uma solução fechada.

$$\beta_{\lambda}^{\mathrm{ridge}} = (X^TX + \lambda I_p)^{-1}X^T y \tag{18}$$

4.4.1. Seleção e ajuste de modelos quimiométricos lineares para predição das variáveis analíticas

Esta etapa do trabalho tem por objetivo identificar as correlações entre dados de espectroscopia Raman e as variáveis caracterizadas *off-line* do conjunto de amostras.

A seleção de variáveis de espectroscopia é uma etapa muito importante em diversos tipos de modelagem, ao ponto de determinar a eficiência do resultado. Uma vez que dados de espectroscopia são compostos por inúmeros dados, a avaliação de combinações entre eles torna-se um processo impraticável em função do tamanho do modelo proposto. Sendo assim, o objetivo é reduzir o número de variáveis independentes selecionadas para aquelas que contêm somente informação correlacionada à variável de interesse, aumentando a seletividade e consequentemente, melhorando a modelagem estatística.

O exemplo mais simples de seleção de variáveis é a seleção de picos em dados de espectroscopia, onde diversos dados podem ser quantificados utilizando combinações entre elementos espectrais que apresentam picos. Neste trabalho será feito uso de uma ferramenta de Otimização Estocástica chamada Colônia de Formigas.

A utilização desta técnica torna viável a seleção de variáveis, utilizando como critério de seleção a capacidade de predição do modelo para a variável de interesse avaliada.

4.5. Validação de modelos

De posse dos modelos quimiométricos ajustados, os dados de espectroscopia foram então utilizados para a predição dos dados de caracterização off-line das amostras de solo, através de modelos com diferentes quantidades de variáveis de entrada (comprimentos de onda).

Como método para avaliar se os modelos ajustados são robustos, o conjunto de amostras foi segmentado em subgrupos de calibração e teste. Os dados de calibração foram

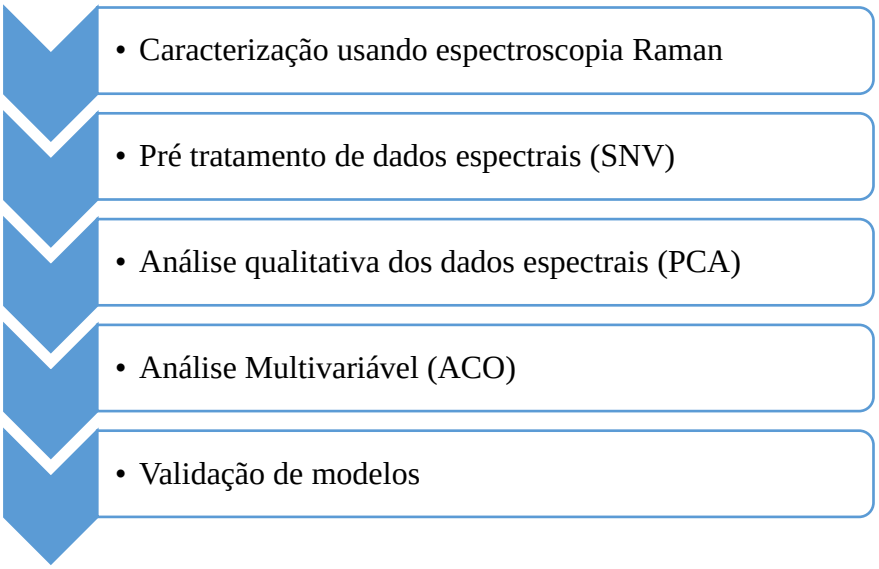
utilizados no ajuste e otimização dos modelos, e o modelo calibrado então é utilizado para predizer os dados de teste, nunca antes vistos pelo modelo. Se as métricas de teste forem comparáveis com as métricas de calibração, então o modelo pode ser considerado robusto. A separação dos subgrupos seguiu a metodologia de amostragem sistemática modificada, onde os dados são ordenados de acordo com a resposta da variável de saída desejada, e cada *k-ésima* amostra é selecionada para o grupo de teste. O tamanho do passo entre *k*'s é definido de acordo com a proporção desejada. O algoritmo modificado utilizado garante que as amostras com menor e maior valor para a variável de saída sejam selecionadas para o grupo de calibração, evitando que o modelo ajustado necessite fazer extrapolações.

Para poder aplicar a técnica de analise multivariável nos dados de espectroscopia, inicialmente o conjunto de amostras foi segmentado em dois grupos, um utilizado para calibração dos modelos, e outro para testes dos modelos calibrados. A segmentação do conjunto de amostras foi realizada seguindo os preceitos da amostragem sistemática modificada, descrita no item 3.6, separando os dados originais em subgrupos de calibração e teste com, respectivamente, 70% e 30% dos dados totais (RANZAN, 2014a).

4.6. Visão geral da metodologia

A metodologia proposta para este trabalho é apresentada de forma esquemática através do diagrama presente na Figura 5.

Figura 5: Representação da metodologia proposta na forma de diagrama.



Fonte: Autoria própria.

5. RESULTADOS E DISCUSSÃO

Para fins de simplificação e esclarecimento, os resultados deste trabalho são apresentados de forma segmentada, em função da metodologia desenvolvida, na seguinte ordem: (i) Amostras e caracterização laboratorial; (ii) Dados de Espectroscopia Raman; (iii) Análise de Componentes Principais (PCA); (iv) Análise Multivariável– ACO.

5.1. Amostras e Caracterização Laboratorial

De posse do conjunto de amostras de solo, foi realizada criação de um banco de dados contendo os dados apresentados pelos laudos do Labsolos.

Cada amostra foi previamente caracterizada pelo Labsolos, sendo que o tempo médio estimado para a obtenção de cada laudo foi de um mês, podendo ser realizadas de maneira simultânea ou não.

Os dados de caracterização laboratorial das amostras de solos foram sintetizados na Tabela 1.

Cada uma das dez amostras de solo, utilizada neste estudo de caso, foi caracterizada com relação ao seu conteúdo de Argila (%), pH, índice SMP, Fósforo, Potássio, Matéria Organiza, Alumínio, Cálcio, Magnésio, H+AL e CTC. Cada uma destas variáveis foi investigada de acordo com seu potencial de predição através de modelos lineares em função de dados de espectroscopia Raman.

Tabela 1: Caracterização quantitativa laboratorial do conjunto amostral, fornecido pelo Labsolos FURG/SAP, sendo 1 cmolc igual a 10 mg de hidrogênio.

Amostra	Argila (%)	pH	Índice SMP	Fósforo (mg/L)	Potássio (mg/L)	Matéria Orgânica (%)	Alumínio (cmolc/L)	Cálcio (cmolc/L)	Magnésio (cmolc/L)	H+Al (cmolc/L)	CTC (cmolc/L)
1	6	5,2	6,2	9,3	22,3	1,8	0,2	3,4	1,7	3,5	8,6
2	15	6,8	6,8	185,6	127,2	5,3	0,1	24	4	1,7	28,4
3	10	5,4	6,1	37,9	51	2,8	0,1	3,4	5,9	3,9	9,5
4	5	5,3	6,7	61	12,4	0,7	0,1	1	0,6	2	4,6
5	15	4,3	4,9	10,7	50,4	8,3	1,4	7,4	4,2	15,4	13,1
6	7	5,5	6	6,6	50	5,2	0,1	10,2	2,7	4,3	17,3
7	8	7,8	7,2	290,7	174	3,8	0,1	11,3	6,9	1,1	19,7
8	47	5,7	5,9	26,5	16,2	2,2	0,2	6,2	3,1	4,9	14,2
9	12	4,7	5,1	12,2	12,8	7,1	2	1,6	0,7	12,2	4,3
10	12	5,7	6,2	25,2	20,7	3,3	0,1	7,8	6,4	3,5	14,3

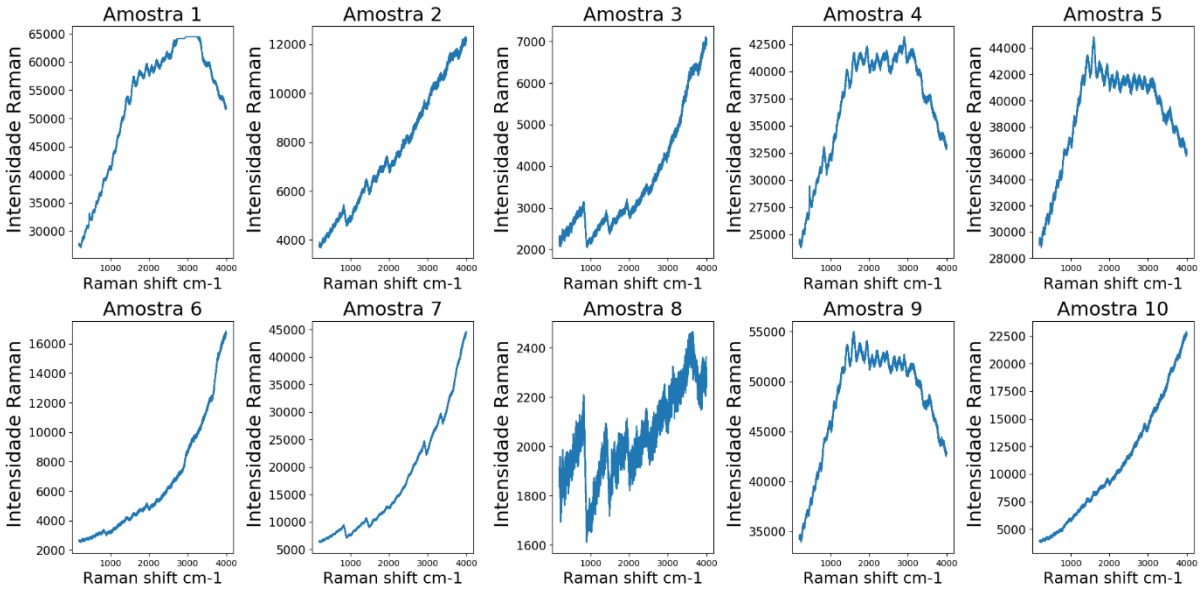
Fonte: Autoria própria.

5.2. Dados de Espectroscopia Raman

De posse do banco de dados das amostras caracterizadas analiticamente, deu-se início à caracterização das mesmas através da técnica de espectroscopia Raman. Foi realizada a coleta dos dados de espectroscopia de cada uma das amostras que compõe o conjunto amostral. Estes dados foram obtidos seguindo a metodologia descrita no item 4.2.

A Figura 6 apresenta os dados de espectroscopia Raman para as 10 amostras de solo analisadas. Nesta figura, uma vez que para cada amostra foram coletados quatro espectros, é apresentado o espectro Raman médio de cada amostra. Esta média aritmética entre os espectros é calculada e apresentada unicamente nesta figura com fins qualitativos, uma vez que a apresentação de todos os espectros coletados não agrega valor à informação visual.

Figura 6: Representação gráfica dos dados de espectroscopia Raman médio obtidos para o conjunto amostral em estudo.



Fonte: Autoria própria.

A Figura 6 permite uma análise primaria das características do conjunto amostral. Através dessa figura pôde-se identificar que o conjunto de amostras é significativamente distinto, uma vez que os espectros não apresentaram forma uniforme. Alguns conjuntos de amostras apresentam um pequeno grau de equivalência, entretanto, de forma geral, os dados de espectroscopia revelam uma grande disparidade no conjunto amostral.

A disparidade nos conjuntos de dados espectrais já era esperada. Técnicas de espectroscopia são, em sua grande maioria, técnicas altamente sensíveis à composição da

matriz de compostos que compõe a amostra. Esta é uma das suas principais vantagens e, simultaneamente, desvantagem deste tipo de análise.

Associado a este fato, solo é um meio significativamente heterogêneo e anisotrópico. Sendo assim, mesmo amostras coletadas em uma mesma propriedade apresentam elevada diferença com relação à composição química. Portanto, amostras coletadas ao longo da área de uma cidade, com grande extensão de terras como Santo Antônio da Patrulha, devem apresentar uma grande diferença em relação aos componentes (confirmado pelos dados contidos na Tabela 1) e com consequência, grande diferença com relação aos seus dados espectrais.

O conjunto de dados amostrais avaliado neste estudo não foi segmentado em função do posicionamento geográfico, de forma que não possuímos informações que possam categorizar o conjunto amostral em função da posição geográfica de coleta das amostras. Tais informações poderiam ser de grande valia para a interpretação qualitativa dos dados espectrais, entretanto, não são o foco deste estudo.

Assim como o conjunto de dados quantitativos, apresentados na forma da Tabela 1, os dados espectrais também foram dispostos na forma de tabela, onde cada linha representa uma das quatro medidas espectrais, coletada para cada amostra, totalizando assim 40 linhas.

Nas colunas desta tabela estão dispostos os comprimentos de onda que foram avaliados pelo equipamento, dentro do range espectral descrito na metodologia. Esta tabela composta por 40 linhas e 7092 colunas compõe o conjunto de dados espectrais.

De posse dos dois conjuntos de dados, qualitativo e espectral, foi iniciado o processo de caracterização qualitativa e quantitativa das amostras de solo.

5.3. Análise de Componentes Principais (PCA)

A partir da análise visual dos dados de espectroscopia do conjunto amostral, apresentado na Figura 6, associado à falta de caracterização qualitativa da origem de cada amostra, optou-se pela realização da Análise de Componentes Principais (PCA), afim de quantificar a heterogeneidade do conjunto amostral.

Para isso, os dados contidos na matriz espectral foram normalizados e analisados através da matriz de *loadings*, obtida através da função PCA, que já é uma função existente no pacote utilizado em Python.

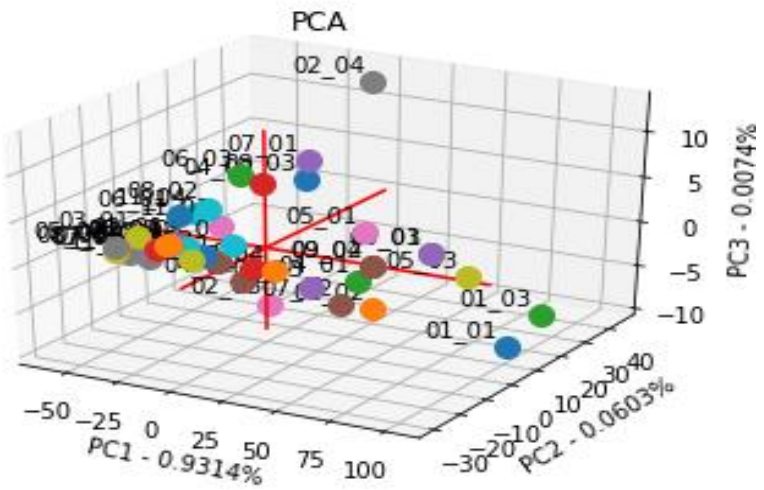
A análise de componentes principais do conjunto de dados espectrais mostrou que 93,14% da variância dos dados está contida no primeiro componente principal, e que 99,91% da variância dos dados pode ser acumulada usando unicamente 3 componentes principais.

A Figura 7 apresenta o posicionamento de cada espectro, para as 10 amostras, em função dos três principais componentes principais. Nesta figura, cada ponto representa um dos espectros coletados, onde os dois primeiros dígitos representam a amostra (variando de 01 a 10), enquanto dos dois últimos algarismos representa o espectro coletado de cada amostra (variando de 01 à 04). Sendo assim, o ponto indicado por 02_04 faz referência ao quarto espectro coletado da segunda amostra.

Na Figura 7, pode ser observado que não existiu uma segmentação significativa entre as amostras. Neste tipo de análise objetiva-se a identificação de conglomerados de amostras, indicando que estas amostras apresentam características espectrais equivalentes. Na figura em questão, não existiu a formação destes conglomerados, mas sim, foi formado um único grupo coeso de pontos, com alguns elementos mais externos.

É interessante salientar que, mesmo espectros coletados de uma mesma amostra de solo, não aparecem em agrupamentos coesos, indicativo de que mesmo em uma amostra, a elevada sensibilidade da técnica de análise espectral, foi capaz de identificar a grande heterogeneidade de cada amostra.

Figura 7: Análise de PCA do conjunto bruto de dados analisados.

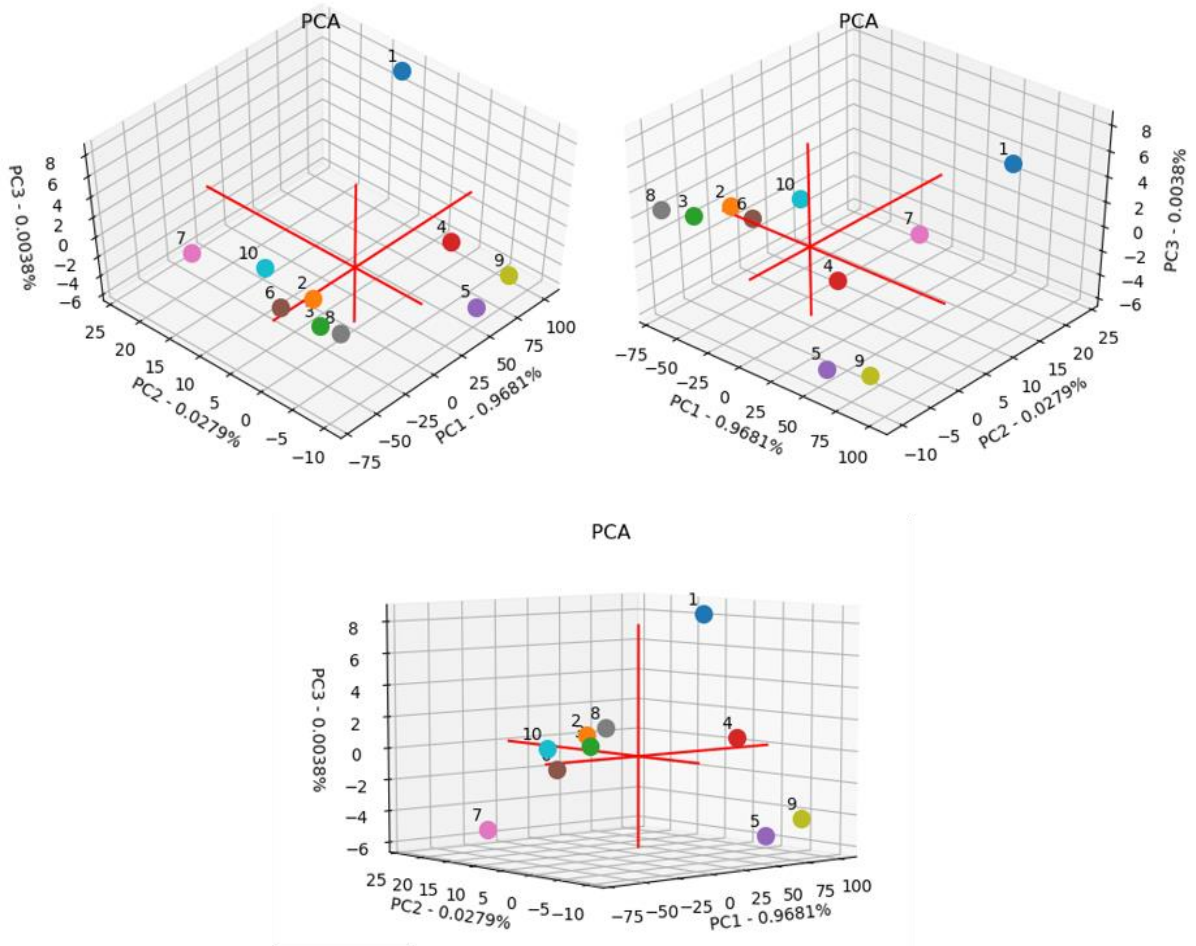


Fonte: Autoria própria.

Uma vez que não houve o agrupamento das medidas espectrais de cada amostra, foi realizada a análise de PCA para o conjunto de dados espectrais médios, afim de identificar similaridades genéricas entre as amostras. A Figura 8 apresenta o posicionamento de cada amostra, em função dos três primeiros componentes principais, para os espectros médio de cada amostra.

Na Figura 8 é possível identificar um grau de equivalência entre algumas amostras, com a formação de dois agrupamentos. O primeiro agrupamento é formado pelas amostras 02, 03, 06, 08 e 10, enquanto o segundo agrupamento é formado pelas amostras 04, 05, e 09. Tanto a amostra 01, quanto a amostra 07, se encontram isoladas das demais. Comparando o resultado apresentado na Figura 8, com os dados espectrais apresentados na Figura 6, é possível identificar que existe uma correlação entre a forma apresentada pelos espectros e a formação destes agrupamentos entre as amostras.

Figura 8: Análise de PCA da média dos espectros para o conjunto de dados analisados.



Fonte: Autoria própria.

5.4. Análise Multivariável (ACO)

Após a análise qualitativa dos dados espectrais, através dos espectros e da análise em componentes principais dos mesmos, deu-se início à análise multivariável dos conjuntos de

dados espectrais associado ao conjunto de dados quantitativos, com o objetivo de determinar a capacidade potencial de predizer os dados analíticos a partir dos dados espectrais.

Nesta etapa, foram gerados modelos polinomiais multivariáveis, com diferente quantidade de variáveis de entrada (de 2 até 4 variáveis). As variáveis de entrada destes modelos são os componentes espectrais da análise Raman, entretanto, uma vez que cada espectro é composto por milhares elementos espectrais é necessário realizar um processo de seleção capaz de determinar quais conjuntos de elementos são mais indicados para predizer a característica analítica desejada.

A ferramenta escolhida para realização desta tarefa foi aquela apresentada no item 4.3. De forma geral, esta metodologia busca combinações de elementos espectrais, através de modelos polinomiais e lineares nos parâmetros (ajustados por regressão Ridge).

O algoritmo empregado na busca dos conjuntos de elementos espectrais foi executado, individualmente, para cada variável de interesse (características das amostras de solo contidas na Tabela 1). Após cada execução, o algoritmo retornou o tamanho de modelo polinomial mais adequado para mensurar cada propriedade, os elementos espectrais selecionados, os parâmetros ajustados e a qualidade da predição, expressa em RMSE, MAPE e R^2.

Previamente à execução do algoritmo, foi necessária a definição de alguns parâmetros de entrada do algoritmo.

Neste trabalho, os valores definidos para os parâmetros do algoritmo foram: tamanho de modelos de 2 a 4 variáveis; a métrica otimizada durante a rodagem do algoritmo – RMSE; número de formigas - 125; número de iterações – 100; número de rodagens (quantas vezes o algoritmo é executado do início) – 10; Taxa de evaporação de feromônio – 10%; Tipo de modelo linear ajustado – Ridge Regression; Valor inicial da trilha de feromônios para cada variável – 1000; Numerador proporcional para adição de feromônio – 1000.

A comparação entre a capacidade preditiva dos modelos ajustados foi baseada em uma ponderação entre as métricas R^2, MAPE e RMSE para o subgrupo de calibração e de teste.

Uma vez que metodologia de otimização empregada pelo algoritmo utilizado trata da execução sequencial da seleção de conjuntos de elementos e ajustes de modelos, ponderada por uma variável chamada trilha de feromônios, o tempo de execução do algoritmo é diretamente proporcional à quantidade de iterações e de formigas.

A Tabela 2 apresenta os resultados obtidos nessa prospecção primária da capacidade preditiva dos modelos. Nesta tabela estão apresentados os melhores resultados de modelos obtidos para cada uma das 11 variáveis de interesse, selecionados com base no menor valor da média entre o RMSE de calibração e teste.

Com os resultados apresentados na Tabela 2, pode-se concluir que o ajuste de modelos lineares, com base em medidas de Raman para amostras de solo sem tratamento, não conseguiu predizer todas as variáveis de interesse de maneira satisfatória. Nos casos em que o R^2 é inferior a 0, está significado que um modelo constante, igual a média das saídas, é mais preciso que os modelos calibrados.

Para a predição de pH e do Índice SMP, os erros percentuais absolutos médios foram consideravelmente baixos (menores que 11%), porém, uma vez que os valores de saída têm baixa amplitude, as métricas de R^2 ainda assim apresentaram valores baixos.

Tabela 2: Melhores resultados obtidos em cada parâmetro na prospecção primária da capacidade preditiva dos modelos.

Variável de Interesse	Tamanho do Modelo	MAPEcal (%)	MAPEteste (%)	R^2cal	R^2teste	RMSEcal	RMSEteste	Variáveis Selecionadas
Alumínio	2	181,510	107,667	0,535	0,777	0,477	0,260	(911.66, 2865.52)
Argila	4	80,887	59,816	0,146	-2,557	13,287	6,668	(2997.35, 3195.09, 3608.26, 3624.34)
CTC	2	37,273	48,487	0,525	0,000	5,434	5,138	(992.05, 3345.68)
Cálcio	3	109,822	54,882	0,361	0,420	6,096	2,739	(1955.04, 2973.23, 3339.78)
Fósforo	3	310,814	181,551	0,084	0,294	96,578	59,594	(1566.52, 1569.74, 3325.85)
H+Al	3	76,396	129,732	0,527	-1,049	3,271	5,687	(233.22, 1912.71, 2836.58)
Indice_SMP	3	8,924	6,223	0,207	0,381	0,655	0,424	(906.30, 959.36, 2905.18)
Magnésio	3	90,613	49,072	0,334	0,419	1,800	1,566	(202.68, 243.94, 2292.65)
Matéria Orgânica	3	98,697	44,596	0,049	0,140	2,349	1,837	(245.55, 2864.45, 3979.10)
Potássio	4	115,710	154,965	0,160	-0,229	51,248	49,549	(2175.83, 2306.59, 3945.34, 3946.41)
pH	2	10,844	7,095	0,329	0,444	0,877	0,522	(902.55, 1462.02, 3174.73, 3346.75)

Fonte: Autoria própria.

Apesar do baixo valor de R^2 obtido, optou-se por ampliar a busca por modelos preditivos unicamente para esta variável (pH), a fim de verificar a possibilidade de incrementar a capacidade preditiva do modelo através da mudança dos parâmetros do algoritmo, como o número de iterações e formigas, aumentado, consequentemente, o tempo de execução do algoritmo.

O aumento do número de iterações e formigas visa aumentar a possibilidade de o algoritmo convergir para um modelo (seleção de elementos espectrais e ajuste de parâmetros) com capacidade preditiva mais satisfatória do valor de pH do solo. Uma das possíveis causas para a baixa capacidade preditiva dos modelos ajustados pode ser associada à ocorrência de mínimos locais. Esta característica é inerente à natureza do problema de otimização, uma vez que o algoritmo busca selecionar de dois a quatro elementos espectrais em um conjunto de

7092 elementos disponíveis, possuindo unicamente 40 espectros (4 espectros coletados por amostra em um conjunto composto por 10 amostras) para fornecimento de informação útil.

A Tabela 3 apresenta o melhor modelo ajustado para a predição do pH, após a execução do algoritmo ACO utilizando um número maior de iterações (200), rodagens (50) e formigas (400).

Tabela 3: Melhor modelo ajustado para caracterização do parâmetro de pH, obtido com 200 iterações e 400 formigas.

Tamanho do Modelo	MAPEcal (%)	MAPEteste (%)	R^2cal	R^2teste	RMSEcal	RMSEteste	Variáveis Selecionadas
2	11,507	6,876	0,316	0,557	0,874	0,426	(902.55, 2913.21)

Fonte: Autoria própria.

O resultado apresentado na Tabela 3 mostra que, apesar do incremento na busca do algoritmo de otimização, o resultado apresentado pelo melhor modelo foi similar na caracterização do pH das amostras de solo, corroborando a suposição de que o tamanho do problema de otimização pode estar inviabilizando a obtenção de conclusões reais sobre a capacidade preditiva da metodologia.

Existe a possibilidade, não explorada nesta primeira análise, de que a correlação entre as variáveis de interesse e os espectros Raman não seja linear. Se este for o caso, o ajuste de modelos puramente lineares nunca seria capaz de predizer de maneira satisfatória as variáveis de interesse.

De forma geral, os resultados, sejam preliminares ou com uma busca mais elaborada, demonstram que a utilização de espectroscopia Raman para a caracterização de amostras de solo não tratadas não apresentou capacidade preditiva satisfatória, obtendo índices de correlação quadrado na etapa de validação de no máximo 0.8, e erros absolutos percentuais médios muito elevados.

Estes resultados não inviabilizam a utilização da espectroscopia Raman com a finalidade de caracterizar amostras de solo, entretanto, novos estudos devem ser realizados a fim de melhorar a metodologia de preparação de amostra (uma vez que neste trabalho optamos por não manipular a amostra de solo). Além deste, novas execuções do algoritmo de otimização devem ser realizadas, aumentando a quantidade de iterações e formigas, e, preferencialmente, o ajuste de modelos não lineares.

Uma vez que a natureza combinatória do problema de otimização é significativamente elevado, é necessário fornecer ao otimizador a maior quantidade de informações possíveis,

sendo assim, o conjunto amostral deve ser aumentado, através da aquisição de novas amostras de solo.

6. CONSIDERAÇÕES FINAIS

A caracterização utilizando a espectroscopia Raman foi a primeira etapa no trabalho na qual se pôde agrupar as amostras em grupos com características semelhantes, o que poderia ser conectado, se em posse da localização geográfica de onde foram retiradas as amostras, com a geolocalização das mesmas.

Na análise de componentes principais (PCA), também se notou a formação de dois grupos, que coincidiram com os grupos formados na caraterização utilizando a espectroscopia Raman, corroborando o que foi analisado anteriormente. Na PCA, o primeiro componente principal contém cerca de 93% da variância dos dados e cerca de 99,9% da variância dos dados pode ser acumulada usando 3 componentes principais.

Na etapa de análise multivariável utilizando o algoritmo colônia de formigas (ACO), a otimização e ajuste de modelos lineares com uso dos dados de espectroscopia Raman não foi satisfatório para predizer as variáveis de interesse. Os erros percentuais de predição, para a maioria das características, foram muito elevados, e os coeficientes de determinação, baixos. Para duas propriedades, pH e Índice SMP, os erros de predição dos modelos foram baixos, mas ainda assim não devem ser considerados como satisfatórios. Vale apontar que para o pH, por ser uma transformação logarítmica, pequenos erros percentuais já podem apontar grandes diferenças. O incremento nos parâmetros de busca, aumentando o número de iterações, rodagens e formigas, não melhorou de forma expressiva os resultados. O pequeno número de amostras, perante a vasta quantidade de variáveis de entrada, pode ter influenciado negativamente na qualidade da busca do algoritmo. Outros fatores a serem considerados são as possiblidades de: (i) a informação contida nos espectros Raman não ter correlação com as variáveis de interesse, ou ainda (ii) a correlação entre as variáveis de interesse e os espectros não ser linear.

Uma análise destes fatos revela que a metodologia proposta não apresentou resultados satisfatórios na predição das variáveis de interesse, indicando que a metodologia de análise empregada necessita ser aperfeiçoada, como, por exemplo, o manuseio das amostras previamente à coleta dos espectros Raman, eliminando possíveis interferentes, homogeneizando o meio e tornando os dados mais representativos.

A espectroscopia Raman neste trabalho não se mostrou satisfatória para a predição de características físico químicas de amostras de solos. Mesmo assim, esta é uma área que merece ser estudada mais a fundo, incluindo o uso de outros tipos de espectroscopia (NIR, IR, UV), além de um pré-tratamento das amostras. A diferença de tempo que fora investido para a

realização da caracterização online (espectros coletados em minutos) e das caracterizações em laboratório é muito relevante e necessário nos dias atuais em que tempo e praticidade estão intimamente conectados com valorização econômica.

CRONOGRAMA DE ATIVIDADES

Para o desenvolvimento do trabalho, elaborou-se o seguinte cronograma de atividades:

Atividades	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Entrega do tema e justificativa do trabalho										
Entrega dos objetivos e resultados esperados										
Obtenção das amostras de solo										
Entrega da metodologia e cronograma atividades										
Análise online das amostras										
Apresentação parcial										
Entrega da revisão										
Entrega da introdução, resumo e abstract										
Entrega da versão final e definição da banca										
Defesa do PCC										
Pré-tratamento de dados										
Análise qualitativa de dados (PCA)										
Análise Multivariável (ACO+OLS)										
Apresentação Parcial PCC II										
Validação dos modelos										
Defesa do PCC II										

REFERÊNCIAS BIBLIOGRÁFICAS

AGAMI REDDY, T. Applied Data Analysis and Modeling for Energy Engineers and Scientists. **Springer**, v. 1, p. 1689–1699, 2011.

ALLEGRINI, F.; OLIVIERI, A. C. A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy/partial least-squares analysis. **Analytica Chimica Acta**, v. 699, n. 1, p. 18–25, 2011.

BALASTREIRE, L. A. O estado-da-arte da agricultura de precisão no Brasil. p. 224, 2000.

BOSQUE-SENDRA, J. M.; CUADROS-RODRÍGUEZ, L.; RUIZ-SAMBLÁS, C.; DE LA MATA, A. P. **Combining chromatography and chemometrics for the characterization and authentication of fats and oils from triacylglycerol compositional data-A reviewAnalytica Chimica Acta**, 2012. .

BRERETON, R. G. **Chemometrics for Pattern Recognition**. [s.l: s.n.]

CAMACHO, J.; PICÓ, J.; FERRER, A. Data understanding with PCA: Structural and Variance Information plots. **Chemometrics and Intelligent Laboratory Systems**, v. 100, n. 1, p. 48–56, 2010.

CLARK, R. J. H.; DINES, T. J. Resonance Raman Spectroscopy, and Its Application to Inorganic Chemistry. **Angew. Chem. Int. Ed. Engl.**, v. 25, p. 131–158, 1986.

DE MYTTENAERE, A.; GOLDEN, B.; LE GRAND, B.; ROSSI, F. Mean Absolute Percentage Error for regression models. **Neurocomputing**, 2016.

DOUGLAS A. SKOOG, F. JAMES HOLLER, S. R. C. **Principle of Instrumental Analysis**. [s.l: s.n.]v. 53

EVERITT, B. S. **Cambridge Dictionary of StatisticsWest Nyack**, 2002. . Disponível em: <http://site.ebrary.com/lib/phoenixcollege/Doc?id=10070000&page=3>.

FARIA, D. L. a. Entenda o que é espectroscopia Raman. **Quimica Viva - CRQ-IV**, p. 4, 2011. Disponível em: <http://www.crq4.org.br/default.php?p=texto.php&c=quimica_viva>.

FEARN, T. Ridge Regression. **NIR news**, v. 24, n. 3, p. 18–19, 2013. Disponível em: <http://journals.sagepub.com/doi/10.1255/nirn.1365>.

GOSS, S.; ARON, S.; DENEUBOURG, J. L.; PASTEELS, J. M. Self-organized shortcuts in the Argentine ant. **Naturwissenschaften**, v. 76, n. 12, p. 579–581, 1989.

KÄMPF, N.; MARQUES, J. J.; CURI, N. Mineralogia de solos brasileiros. **Pedologia: Fundamentos**, p. 82–145, 2012.

LOPES, A. S. Acidez Do Solo. 1991.

MOLIN, J. P. Agricultura de precisão: números do mercado brasileiro. **Agricultura de Precisão: boletim técnico 3**, p. 7, 2017. Disponível em:

<http://www.agriculturadeprecisao.org.br/publicacoes/categoria/3>.

NASCIMENTO, R. S.; FROES, R. E. S.; E SILVA, N. O. C.; NAVEIRA, R. L. P.; MENDES, D. B. C.; NETO, W. B.; SILVA, J. B. B. Comparison between ordinary least squares regression and weighted least squares regression in the calibration of metals present in human milk determined by ICP-OES. **Talanta**, v. 80, n. 3, p. 1102–1109, 2010.

PEREIRA, H. S. Fósforo e potássio exigem manejos diferenciados. p. 43–46, 2009.

PINHEIRO, R. AGRICULTURA DE PRECISÃO Estudos de uma tecnologia favorável , na suscitação de melhoras na qualidade de técnicas empregadas no campo por José Paulo Molin Studies of a favorable technology , in a raised of improvements in the. p. 53–71, [s.d.]

RANZAN, C. Desenvolvimento De Modelos Quimiométricos Utilizando O Algoritmo De Otimização Colônia De Formigas. p. 150, 2014a.

RANZAN, L. Estudo da viabilidade do uso de espectroscopia por fluorescência 2D para quantificar teor de enxofre em óleo diesel. p. 94, 2014b. Disponível em: <http://hdl.handle.net/10183/116717>.

RESENDE, Á. V; SHIRATSUCHI, L. S.; COELHO, A. M.; CORAZZA, E. J.; MARINA, F.; INAMASU, R. Y.; BERNARDI, A. C. C.; BASSOI, L. H.; NAIME, J. M. Agricultura de Precisão no Brasil: Avanços, Dificuldades e Impactos no Manejo e Conservação do Solo, Segurança Alimentar e Sustentabilidade. **XVIII REUNIÃO BRASILEIRA DE MANEJO E CONSERVAÇÃO DO SOLO E DA ÁGUA Novos Caminhos para Agricultura Conservacionista no Brasil**, 1999.

RODRIGUES, A. D. G.; GALZERANI, J. C. Espectroscopias de infravermelho, Raman e de fotoluminescência: potencialidades e complementaridades. **Revista Brasileira de Ensino de Física**, v. 34, n. 4, 2012. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1806-11172012000400009&lng=pt&tlng=pt>.

SALLES, R. C. Interpretação da análise de solo 1. [s.d.]

SETTLE, F. A. Voltammetric Techniques BT - Handbook of instrumental techniques for analytical chemistry. In: **Handbook of instrumental techniques for analytical chemistry**. [s.l: s.n.]p. 995.

TEDESCO, M. J. **Analysis of soil, plant and other materials**, 1995. .

WEHRENS, R. **Chemometrics with R: multivariate data analysis in the natural sciences and life sciences**. [s.l: s.n.]

ŽUPERL, Š.; FORNASARO, S.; NOVIČ, M.; PASSAMONTI, S. Experimental determination and prediction of bilitranslocase transport activity. **Analytica Chimica Acta**, 2011.