ORGANIZAÇÃO DE UMA ESTRUTURA DE PROCESSAMENTO PARALELO BASEADO EM APACHE SPARK PARA GERENCIAMENTO DE RISCO AGRÍCOLA

M. F. L. Pereira[1,2,*], G. M. Alves[1,3,4], J. M. G. Beraldo[1,5], P. E. Cruvinel[1,3]

[1] *Embrapa Instrumentação Agropecuária, Rua XV de Novembro, 1452, 13560-970, São Carlos, SP*
[2] *Universidade Federal de Mato Grosso, Av. Fernando C. da Costa, 2367, 78060-900, Cuiabá, MT*
[3] *Universidade Federal de São Carlos, Rod. Washington Luís km 235, 13565-905, São Carlos, SP*
[4] *Instituto Federal de São Paulo, Av. Marginal, 585, 13871-298, São João da Boa Vista, SP*
[5] *Instituto Federal de São Paulo, Rua Stéfano D´avassi, 625, 15991-502 – Matão, SP*
[*] *Autor correspondente, e-mail: mauricio@ic.ufmt.br*

Resumo: A grande disponibilidade de dados produzidos atualmente por estações agrometeorológicas, satélites, VANTs, máquinas agrícolas, dentre outros equipamentos possibilita a exploração e organização de técnicas computacionais para a obtenção de informação útil ao produtor, especialmente no que tange aos riscos agrícolas. Como o volume de dados gerado é, em geral imenso, isso torna inviável a execução de tais algoritmos em máquinas convencionais, tornando cada vez mais necessário o uso de arquiteturas paralelas. O objetivo deste trabalho é apresentar a organização de um modelo paralelo baseado no *framework* Apache Spark para gerenciamento de riscos agrícolas, abordando inicialmente os fatores de risco relacionados à indicadores físicos e químicos da qualidade do solo. Para isso, apresenta-se um algoritmo paralelo que, a partir da entrada de imagens tomográficas de solos agrícolas distribui essas imagens e as tarefas de reconhecimento de poros, de densidade e de compactação dos solos entre diversos nós. A modelagem do algoritmo paralelo proposto possibilitou novos caminhos para ganhos de desempenho e redução do tempo necessário para geração de mapas de riscos baseados na qualidade de solos agrícolas.

Palavras-chave: processamento paralelo, apache spark, imagem tomográfica, qualidade de solo

ORGANIZATION OF AN APACHE SPARK-BASED PARALLEL PROCESSING STRUCTURE FOR AGRICULTURAL RISK MANAGEMENT

Abstract: Actually, a large volume of data is being produced by agrometeorological stations, satellites, UAVs, agricultural machines, among other equipment's, and this demands the organization of new computational techniques to obtain useful information for farmers from such data, especially regarding to agricultural risks. Therefore, the data volume is immense, and it is almost impossible to execute them based on algorithm operating in conventional architectures. This paper presents the organization of a parallel model based on the Apache Spark framework for agricultural risk management, initially addressing risk factors related to soil quality. Based in such concept a parallel algorithm is proposed to distribute tomographic images of soil to several nodes, as well as, to ask them for not only pore recognition, but also for soil density and soil compaction measurements from those images. Besides, the developed parallel algorithm allows to obtain a better performance gains and reduction of the time required to generate the maps of risk for the agricultural soil quality.

Keywords: parallel processing, apache spark, tomographic images, soil quality.

1. Introdução

O agronegócio é um importante pilar da economia brasileira e da economia mundial e nele, como tem ocorrido em diversas áreas, a transformação digital recentemente tem alterado a forma de execução de alguns processos agrícolas, com a inclusão cada vez maior da automação, no que vem sendo chamada há alguns anos de agricultura de precisão e de decisão. Nesse recente cenário, o uso

de máquinas inteligentes guiadas por GPS para plantar, cultivar e colher com precisão está crescendo no país, trazendo com isso maior economia de insumos, ganhos de produtividade e sustentabilidade. Desta forma, essas máquinas atuam como uma mola propulsora e integradora dentro e fora da cadeia de produção. Além das novas tecnologias embarcadas em máquinas agrícolas utilizadas em campo, existe a necessidade de se explorar técnicas que permitam extrair informação útil de grandes volumes de dados gerado por esses equipamentos. Nesse contexto, em especial, é importante desenvolver algoritmos capazes de fazer uso desses dados através de ferramentas computacionais de apoio a decisão, que visem reduzir riscos e consequentemente os prejuízos para o agricultor. Um dos fatores que influência o risco agrícola é a qualidade do solo (QS), dado que ela influencia os índices de produtividade das culturas agrícolas. De forma geral, a Figura 4 apresenta uma adaptação da estrutura para modelos de riscos agrícolas proposto por Cruvinel *et al.* (2017), no qual o "Mapeamento da qualidade do solo" é uma das componentes no cálculo dos riscos agrícolas. Os algoritmos utilizados nessa componente foram modelados nesse trabalho.

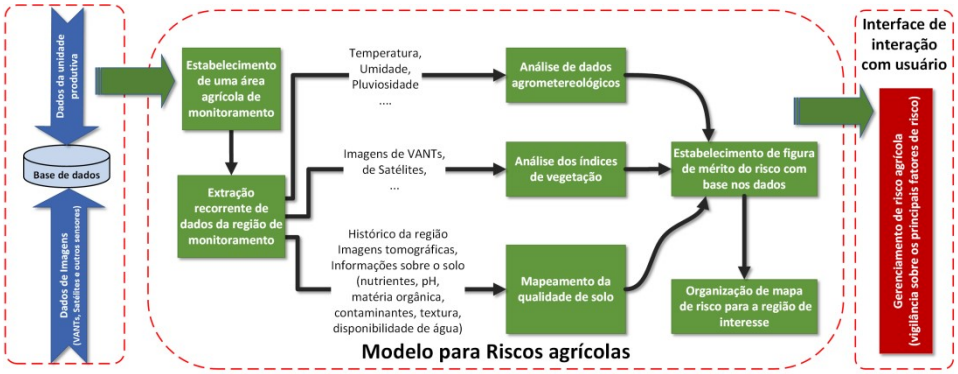

Figura 4. Estrutura para gerenciamento de riscos agrícolas, baseado no uso de geotecnologias e sistemas embarcados de suporte à decisão [Fonte: (CRUVINEL et al., 2017)]

Esse trabalho propõe o desenvolvimento de ferramentas digitais avançadas de apoio ao gerenciamento de riscos agrícolas.

Dado ao grande volume de dados e o custo computacional dos algoritmos, os mesmos serão implementados com uso de técnicas de processamento distribuído com uso do modelo MapReduce (HASHEM et al., 2016), combinados ao uso de infraestrutura de alto desempenho, em *clusters* computacionais dedicados ou serviços em nuvem da *Amazon Web Services* (AWS). No desenvolvimento, a programação será baseada no framework para processamento Apache Spark (ZAHARIA et al., 2016) com uso da linguagem Scala e Python para implementação dos modelos. Inicialmente utilizaram-se imagens tomográficas de solos, as quais permitem inferir medidas de porosidade, compactação e densidade do solo, que são importantes para mapeamento da qualidade de solo (ANDREWS; KARLEN; CAMBARDELLA, 2004; COUNCIL, 1993; SNAKIN et al., 1996; SUMNER, 1999).

2. Materiais e Métodos

Dado ao grande volume de dados e o custo computacional dos algoritmos, os mesmos serão implementados com uso de técnicas de processamento distribuído com uso do modelo MapReduce (HASHEM et al., 2016), combinados ao uso de infraestrutura de alto desempenho, em *clusters* computacionais dedicados ou serviços em nuvem da *Amazon Web Services* (AWS). No desenvolvimento, a programação será baseada no framework para processamento Apache Spark (ZAHARIA et al., 2016) com uso da linguagem Scala e Python para implementação dos modelos. Inicialmente utilizaram-se imagens tomográficas de solos, as quais permitem inferir medidas de porosidade, compactação e densidade do solo, que são importantes para mapeamento da qualidade de solo (ANDREWS; KARLEN; CAMBARDELLA, 2004; COUNCIL, 1993; SNAKIN et al., 1996; SUMNER, 1999).

2.1. MapReduce e Apache Spark

Na indústria e na pesquisa científica, por conta da recente facilidade em se gerar e armazenar dados, os volumes destes tem ultrapassado a capacidade de processamento de dados de um único computador, havendo a necessidade de se utilizar arquiteturas computacionais escaláveis e compostas de múltiplos nós processadores (ZAHARIA et al., 2016). O resultado dessa mudança de paradigma tem sido os novos modelos de programação direcionados a particionar a carga de trabalho em diversos processadores (DEAN; GHEMAWAT, 2008). Os primeiros modelos de programação paralela, como o *Message-Passing Interface* (MPI), por serem mais complexos, exigiam dos usuários maior cuidado e tratamento de problemas, tais como sincronização entre os processos, recuperação dos dados em caso de perda, dentre outros. Já os modelos mais recentes como o *MapReduce*, o qual possui uma forma de implementação de algoritmos associada ao modelo, permite lidar com grandes volumes de dados através de algoritmos paralelos baseados nas funções *map* e *reduce* e no sistema de execução subjacente, o qual toma todos os cuidados para paralelização das tarefas através dos cluster, que em geral possui uma grande quantidade de núcleos. Também ficam sob responsabilidade do sistema de execução, a manipulação de falhas em máquinas e a organização da intercomunicação entre os nós do cluster, de forma a fazer um uso otimizado da rede e dos discos (DEAN; GHEMAWAT, 2008; HASHEM et al., 2016).

Ainda que o *MapReduce* diminuísse bastante a complexidade do desenvolvimento de algoritmos paralelos, em 2009 foi iniciado na Universidade de Berkeley o projeto Apache Spark. Tal projeto estabeleceu como principal objetivo a construção de um ambiente unificado para processamento distribuído de dados, cujo modelo de programação viesse a ser semelhante ao modelo *MapReduce*. Entretanto, considerando a importante abstração de dados denominada *Resilient Distributed Data* (RDD). Essa abstração foi considerada como sendo a mais importante, contribuição, uma vez que permite que objetos sejam tolerantes a falhas e particionados através do cluster computacional. Os RDDs podem ser manipulados em paralelo, o que permite o ganho de desempenho nas aplicações. Usuários podem criar um RDD pela aplicação de transformações, sendo que as mais conhecidas são os filtros (*filters*), mapeamentos (*map*) e agrupamentos (*group by*). O Spark torna possível trabalhar com uma maior variedade de tipos de carga de trabalho computacional, que antes de seu desenvolvimento necessitava de diferentes motores de processamento para tratar situações em que se necessitava trabalhar com SQL, ou com *streaming*, ou com aprendizagem de máquina e processamento de grafos. No Spark todas essas situações mencionadas estão embutidas no núcleo de seu motor. Os principais benefícios oferecidos pelo Spark são: (1) Aplicações mais facilmente desenvolvidas através da API unificada oferecida no *framework*; (2) Maior eficiência nas tarefas que demandem processamento combinado, ou seja, um sistema cujo primeiro módulo prepare informações a partir da base de dados e as armazene para um outro motor ou sistema processar essas informações. No Spark todo esse processo pode ser feito de forma integrada, sobre os mesmos dados e muitas vezes em memória. Mais informações a respeito desse *framework* podem ser obtidas em (SPARK, 2019).

2.2. Qualidade do solo (QS)

Os efeitos diferenciados nos atributos do solo, devido ao tipo de preparo, característico de cada sistema de cultivo, são dependentes do trânsito de máquinas, do tipo de equipamento utilizado, do manejo dos resíduos, das condições de umidade do solo no momento do preparo. A relação entre o manejo e a qualidade do solo pode ser avaliada pelo comportamento de indicadores da qualidade do solo que podem ser classificados, de um modo geral, em quatro grupos; visuais, físicos, químicos e biológicos. Embora esta divisão em grupos seja usual, é importante salientar que estes atributos e processos, em sua maioria, são inter-relacionados. Os indicadores visuais podem ser obtidos a partir da interpretação de fotografias obtidas com aeronaves tripuladas ou drones, ou ainda com imagens de satélites, ou através de observações diretas (verdade de campo), como a exposição do subsolo, mudança de cor do solo, escorrimento superficial, resposta da planta, espécies de plantas daninhas predominantes, entre outras.

2.3. Organização do modelo paralelo para avaliação de risco na qualidade do solo

O modelo de integração dos riscos é baseado na metodologia proposta por Cruvinel e colaboradores (CRUVINEL et al., 2011), onde o risco integrado é fruto da intersecção das probabilidades sobre as faixas de valores das variáveis pré-selecionadas.

3. Resultados e Discussão

A partir da necessidade de qualificação do solo como um dos elementos para mapeamento dos riscos agrícolas, se estabeleceu um modelo paralelo que utiliza os indicadores físicos, químicos e biológicos do solo na região de interesse, conforme descrito na Eq. 1. Conforme apresentado na Eq. 1 o resultado da probabilidade da qualidade do solo passou a ser função da intersecção das probabilidades, ou seja, porosidade, compactação, densidade, nutrientes (como por exemplo nitrogênio (N), fósforo (P), potássio (K)), matéria orgânica, pH, contaminantes (metais pesados, como por exemplo o Chumbo (Pb), o Cobre (Cu) e o Manganês (Mg)), textura (argila, silte e areia) e disponibilidade de água. A textura ou granulometria indica as frações do solo. A argila é a que possui maior superfície específica e é de natureza coloidal com alta retenção de cátions e adsorção de fósforo. A fração argila representa a maior parte da fase sólida do solo e é constituída de uma gama variada de minerais que apresentam cargas elétricas negativas responsáveis pela capacidade de troca de cátions (CTC).

$$P(QS) = P(\text{porosidade}) \cap P(\text{compactação}) \cap P(\text{densidade}) \cap P(\text{nutrientes}) \cap P(\text{matéria orgânica}) \cap P(\text{pH}) \cap P(\text{Contaminantes}) \cap P(\text{textura}) \cap P(\text{disponibilidade de água}) \quad (1)$$

Adicionalmente, a partir de imagens tomográficas do solo obtidas em alta resolução, pode-se extrair informações a respeito da porosidade, densidade e compactação do solo de forma automática, acelerando-se tanto o reconhecimento dessas características, quando o cálculo da figura de mérito para cada área no mapa da região de interesse. Dessa forma, organizou-se a modelagem do algoritmo, mostrada na Figura 5. Nesse contexto foi utilizado o ambiente Apache Spark, para realizar o reconhecimento dessas características a partir de imagens reconstruídas também em ambiente paralelo. Observa-se que os módulos que podem ser paralelizados são mostrados com diversas caixas, ilustrando que o cálculo é realizado de forma distribuída.

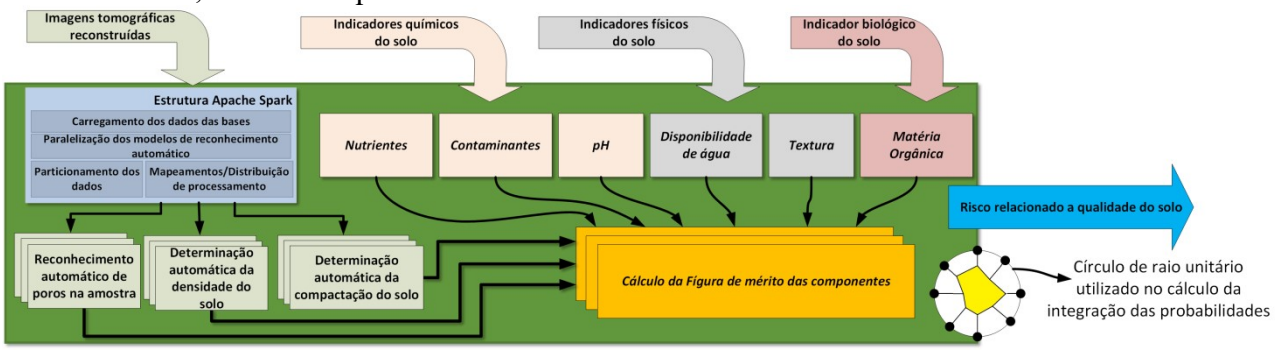

Figura 5. Organização do algoritmo paralelo para determinação do risco relacionado a qualidade do solo, através do reconhecimento automático da porosidade, densidade e compactação através de imagens tomográficas.

Dessa forma, é possível usar os dados de tomografia para avaliação da porosidade, densidade e compactação do solo. Assim, tal como apresentado na Figura 5, tem-se na entrada do algoritmo paralelo imagens tomográficas de solo reconstruídas, que serão lidas no ambiente Apache Spark, assim como os demais dados lidos em laboratório e distribuídas ao longo dos nós na forma de RDD. Aqui, admite-se que se utilizará um cluster com ao menos 4 nós de processadores *multicore*. A partição dos dados se dá de forma que cada nó recebe uma imagem, da qual irá extrair os atributos físicos mencionados. Coordenados pelo algoritmo paralelo, cada nó retorna os atributos físicos e a identificação da imagem analisada. Essas informações fomentam juntamente com os

indicadores da qualidade do solo, quais deverão ser inseridos no sistema através da base de dados, assim como sobre as condições necessárias para o estabelecimento de um mapa de risco da região de interesse. Para cada pixel do mapa da região de interesse, se faz o cálculo da integração de probabilidades de forma paralela, para se obter o resultado.

Apesar de aumentar a complexidade do algoritmo de mapeamento dos riscos, ao se inserir momentos no código para estabelecimento de comunicação entre os nós, o uso do Apache Spark contribui para que se aplique a solução proposta em situações onde se pode ter alta resolução nos mapas de risco. Nessa situação o uso de arquitetura convencional se torna proibitivo dada a demanda de memória e velocidade de processamento.

4. Conclusões

O uso de arquiteturas paralelas de processamento de dados combinadas com a plataforma Apache Spark tem sido bastante utilizado em situações de grandes volumes de dados. Este trabalho apresentou a modelagem de um algoritmo paralelo para a geração de mapas de riscos à qualidade de solo. Os resultados indicam que além do caráter inovador há grande potencial para ganho de desempenho, assim como, para obtenção de mapas de risco considerando as probabilidades de ocorrência das variáveis e sua integração em uma figura de mérito para análise sistêmica.

Agradecimentos

Este trabalho recebe apoio financeiro da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Processo 17/19350-2, via convênio que envolve a IBM Brasil e a Empresa Brasileira de Pesquisa Agropecuária (Embrapa). Os autores também agradecem o apoio institucional da Universidade Federal de Mato Grosso (UFMT) e do Instituto Federal de São Paulo.

Referências

ANDREWS, S. S.; KARLEN, D. L.; CAMBARDELLA, C. A. The Soil Management Assessment Framework: A Quantitative Soil Quality Evaluation Method. . Soil Science Society of America Journal, v. 68, p. 1945–1962, 2004.

COUNCIL, N. R. Soil and water quality: an agenda for agriculture. [s.l.] National Academies Press, 1993.

CRUVINEL, P. E. et al. Avaliação de modelos para estabelecimento de figura de risco de ocorrência da Sigatoka-Negra em bananais. Embrapa Instrumentação-Documentos (INFOTECA-E), p. 28, 2011.

CRUVINEL, P. E. et al. Ferramenta digital avançada para o gerenciamento de riscos agrícolas. Disponível em: <https://bv.fapesp.br/pt/auxilios/100768/ferramenta-digital-avancada-para-o-gerenciamento-de-riscos-agricolas/>.

DEAN, J.; GHEMAWAT, S. MapReduce: Simplified Data Processing on Large Clusters. Commun. ACM, v. 51, n. 1, p. 107–113, 2008.

HASHEM, I. A. T. et al. MapReduce: Review and open challenges. Scientometrics, v. 109, n. 1, p. 389–422, 2016.

SNAKIN, V. V et al. The system of assessment of soil degradation. Soil Technology, v. 8, n. 4, p. 331–343, 1996.

SPARK, A. Apache Spark - Lightning-fast unified analytics engine. Disponível em: <https://spark.apache.org/>. Acesso em: 11 set. 2019.

SUMNER, M. E. Handbook of soil science. [s.l.] CRC press, 1999.

ZAHARIA, M. et al. Apache Spark: A Unified Engine for Big Data Processing. Commun. ACM, v. 59, n. 11, p. 56–65, 2016.