UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE CIÊNCIAS RURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA AGRÍCOLA

Raí Augusto Schwalbert

IMAGENS DE SATÉLITE PARA PREDIÇÃO ESPAÇO-TEMPORAL DA PRODUTIVIDADE DE MILHO E SOJA EM DIFERENTES ESCALAS GEOGRÁFICAS

Santa Maria, RS
2019

Raí Augusto Schwalbert

IMAGENS DE SATÉLITE PARA PREDIÇÃO ESPAÇO-TEMPORAL DA PRODUTIVIDADE DE MILHO E SOJA EM DIFERENTES ESCALAS GEOGRÁFICAS

Tese apresentada ao curso de doutorado em Engenharia Agrícola da Universidade Federal de Santa Maria (UFSM), como requisito parcial para obtenção do grau de **Doutor em Engenharia Agrícola.**

Orientador: Prof. Dr. Telmo Jorge Carneiro Amado

Santa Maria, RS
2019

SCHWALBERT, RAI
IMAGENS DE SATÉLITE PARA PREDIÇÃO ESPAÇO-TEMPORAL DA PRODUTIVIDADE DE MILHO E SOJA EM DIFERENTES ESCALAS GEOGRÁFICAS / RAI SCHWALBERT.- 2019.
95 p.; 30 cm

Orientador: TELMO AMADO
Tese (doutorado) - Universidade Federal de Santa Maria, Centro de Ciências Rurais, Programa de Pós Graduação em Engenharia Agrícola, RS, 2019

1. Imagens de satélite 2. Predição de produtividade 3. Machine Learning 4. Big Data 5. Deep Learning I. AMADO, TELMO II. Título.

Sistema de geração automática de ficha catalográfica da UFSM. Dados fornecidos pelo autor(a). Sob supervisão da Direção da Divisão de Processos Técnicos da Biblioteca Central. Bibliotecária responsável Paula Schoenfeldt Patta CRB 10/1728.

Raí Augusto Schwalbert

IMAGENS DE SATÉLITE PARA PREDIÇÃO ESPAÇO-TEMPORAL DA PRODUTIVIDADE DE MILHO E SOJA EM DIFERENTES ESCALAS GEOGRÁFICAS

Tese apresentada ao curso de doutorado em Engenharia Agrícola da Universidade Federal de Santa Maria (UFSM), como requisito parcial para obtenção do grau de **Doutor em Engenharia Agrícola.**

Aprovado em 10 de Setembro de 2019:

_____
Telmo Jorge Carneiro Amado, Dr. (UFSM)
(Presidente/Orientador)

_____
Ignacio Ciampitti, Dr. (K-State)

_____
Nahuel Peralta, Dr. (Bayer)

_____
Lúcio de Paula Amaral, Dr. (UFSM)

_____
Christian Bredemeier, Dr. (UFRGS)

Santa Maria, RS
2019

RESUMO

IMAGENS DE SATÉLITE PARA PREDIÇÃO ESPAÇO-TEMPORAL DA PRODUTIVIDADE DE MILHO E SOJA EM DIFERENTES ESCALAS GEOGRÁFICAS

AUTOR: Raí Augusto Schwalbert
ORIENTADOR: Telmo Jorge Carneiro Amado

À medida que as questões relacionadas à segurança alimentar global se tornam cada vez mais desafiadores, estimativas confiáveis da produtividade de culturas agrícolas passam a ser mais imperativas do que nunca para a comunidade científica. Atualmente, com a maior facilidade de acesso a dados provenientes de sensores embarcados em satélites, essa fonte de informação tem se tornado muito promissora para o desenvolvimento de modelos de previsão de produtividade de culturas agrícolas. Apesar disso, seu uso ainda é limitado na maioria dos esforços operacionais para monitorar produtividade em diferentes escalas geográficas. De maneira geral, os modelos de previsão de produtividade baseados em imagens de satélite podem ser avaliados considerando três aspectos: i) a acurácia das previsões; ii) a antecedência com que a previsão é realizada em relação à data de colheita; e iii) a escala espacial da unidade de previsão, (e.g. país, estado, município, área agrícola, etc.). Os principais objetivos desse estudo foram: i) desenvolver um modelo de previsão de produtividade com base em imagens de satélite capazes de predizer a produtividade da cultura do milho (no *Corn Belt* dos Estados Unidos) e da soja (no estado do Rio Grande do Sul- Brasil) nos níveis de condado e município, respectivamente; ii) avaliar o desempenho do modelo após a inclusão de variáveis meteorológicas juntamente aos índices de vegetação derivados de satélite; iii) testar diferentes algoritmos de aprendizado de máquina para prever a produtividade em nível regional; e iv) avaliar a capacidade de generalização dos modelos preditivos desenvolvidos em nível de área agrícola quando aplicados para áreas localizadas em diferentes regiões em relação à onde eles foram parametrizados. Os principais resultados foram: i) modelos preditivos baseados em imagens de satélite e variáveis meteorológicas podem antecipar a produtividade da cultura do milho em até 122 dias (aproximadamente 16 dias antes do primeiro relatório de produtividade de milho em nível estadual da USDA/NASS) com um erro médio absoluto menor que 1 Mg ha^{-1}, e em até 70 dias para a soja com erro médio absoluto de 0,42 Mg ha^{-1}; ii) temperatura do ar, temperatura da superfície do dossel e deficit de pressão de vapor melhoraram o desempenho dos modelos em relação aos modelos baseados apenas em índices de vegetação (NDVI e EVI); iii) o algoritmo *Long Short Term Neural Network* apresentou desempenho superior em comparação com os outros algoritmos testados (e.g. *random forest* e regressão ordinária de mínimos quadrados); iv) os modelos de previsão de produtividade parametrizados em nível de área agrícola apresentaram capacidade de generalização limitada fora dos limites onde foram ajustados, mas as semelhanças nos dados usados para parametrização do modelo podem fornecer diretrizes de como eles podem ser extrapolados. Os resultados apresentados nesse estudo têm potencial para auxiliar agricultores e agentes formuladores de políticas durante o processo de tomada de decisão. Estudos futuros sobre esse tópico devem explorar a fusão de modelos mecanísticos (baseados em processos) com modelos empíricos, a fim de aumentar os limites espaço-temporais de predicabilidade e tornar os modelos menos dependente de dados oriundos de terceiros.

Palavras-chave: Imagens de satélite. Predição de produtividade. Aprendizagem de máquina.

ABSTRACT

SATELLITE IMAGERY FOR SPATIO-TEMPORAL CORN AND SOYBEAN YIELD PREDICTION AT DIFFERENT GEOGRAPHICAL LEVELS

AUTHOR: Raí Augusto Schwalbert
ADVISOR: Telmo Jorge Carneiro Amado

As global food security issues become increasingly challenging, reliable estimates of crop yields are becoming more imperative than ever for the scientific community. Today, with greater ease of accessing remote sensing data from satellite-embedded sensors, this source of information has become very promising for developing crop yield forecast models. Nevertheless, the use of such models is still limited in most operational efforts to monitor crop yield at different geographic scales. In general, satellite-based yield forecast models can be evaluated by considering three aspects: i) the accuracy of the predictions; ii) the date when the yield forecast is released in relation to the crop harvest date; and iii) the spatial scale of the forecasting unit, (e.g. country, state, county, field, etc.). The main objectives of this study were: i) to develop a complete model based on satellite images capable of predicting corn (in the US Corn Belt) and soybean (in the state of Rio Grande do Sul – Brazil) in county and municipality levels, respectively; ii) evaluate the performance of the model after the inclusion of weather variables along with satellite derived vegetation indices; iii) test different machine learning algorithms to predict yield at the regional level; and iv) evaluate the generalization capacity of predictive models developed at field level when applied to fields in different regions from which they were parameterized. The main results were: i) satellite-based predictive models and weather variables can anticipate corn yield by up to 122 days (approximately 16 days prior to the first USDA/NASS state-level corn yield report) with an mean absolute error of less than 1 Mg ha^{-1}, and soybean yield by up to 70 days with an mean absolute error of 0.42 Mg ha^{-1}; ii) air temperature, canopy surface temperature and vapor pressure deficit improved model performance in relation to models based only on vegetation indices (NDVI and EVI); iii) the Long Short Term Memory Neural Network algorithm performed better compared to the other algorithms tested (e.g. random forest and ordinary least squares regression); and iv) the models parameterized at field level presented limited generalization capacity outside the limits where they were adjusted, but similarities in the data distribution used for model parameterization can provide guidance on how they can be extrapolated. The results presented in this study have potential to assist farmers and policy makers in the decision making process. Future studies on this topic should explore the fusion of mechanistic (process-based) with empirical models in order to increase the spatio-temporal limits of predictability and make models less dependent on third party data.

Keywords: Satellite imagery. Yield forecast. Machine learning.

SUMÁRIO

1 APRESENTAÇÃO

A agricultura está passando por uma revolução digital baseada na geração, coleta e interpretação de massiva quantidade de dados. Acesso a informação de qualidade e de maneira antecipada, ou seja, no menor tempo decorrido desde sua coleta, é extremamente interessante no contexto da operação agrícola, com potencial para interferir na tomada de decisão em diferentes esferas, desde a compra dos insumos, manejo da propriedade até comercialização do produto final.

Atualmente, considerável parcela dos dados com potencial prático para influenciar decisões nas operações agrícolas são proveniente de sensores embarcados em satélites. Imagens de satélites possuem ampla aplicação na agricultura (LOBELL, 2013; SAKAMOTO; GITELSON; ARKEBAUER, 2014), em especial na geração de modelos capazes de predizer produtividade em tempo real. Estimativas (pós-colheita) ou previsões (a.k.a predições) (pré-colheita) confiáveis de produtividade podem ser úteis para diversos propósitos, e normalmente sua aplicabilidade está associada à escala em que elas são realizadas. Previsões em nível de áreas agrícolas são particularmente úteis para entender como a produtividade das culturas responde a fatores ambientais e de manejo (LOBELL, 2013; PERALTA et al., 2016), permitindo o uso mais eficiente de recursos como água e fertilizantes. Ao passo que, previsões em domínios maiores (municípios, estados ou países) são úteis para questões envolvendo, políticas governamentais, segurança alimentar, logística e transporte da produção agrícola (SAKAMOTO; GITELSON; ARKEBAUER, 2014), especialmente em países como o Brasil, que desempenha papel importante no mercado internacional de grãos.

Previsões de produtividades baseadas em técnicas de sensoriamento remoto têm sido do interesse de pesquisadores durante muitos anos, inicialmente com foco em escalas regionais (DIRIENZO; FACKLER; GOODWIN, 2000; LOPRESTI; DI BELLA; DEGIOANNI, 2015; MACDONALD; HALL, 1980; SIBLEY et al., 2014), principalmente porque no passado havia acesso limitado a dados com alta resolução espacial, e mais recentemente, à nível de área agrícola (JIN et al., 2017; JIN; AZZARI; LOBELL, 2017; LOBELL et al., 2015; PERALTA et al., 2016). Entre os principais fatores que propiciaram esse maior nível de detalhamento nos estudos atuais destacam-se: i) o lançamento de satélites capazes de adquirir imagens com alta resolução espacial e temporal, incluindo satélites públicos como o Sentinel-2 (DRUSCH et al., 2012), e privados como RapidEye e Skysat, ii) a disponibilização gratuita de imagens de satélites de instituições públicas, como a NASA e a

ESA, e algumas instituições privadas (AZZARI; JAIN; LOBELL, 2017), e iii) o desenvolvimento e aperfeiçoamento de algoritmos e plataformas de processamento de dados, como por exemplo, o Google Earth Engine (GEE) (GORELICK et al., 2017).

1.2 REFERENCIAL TEÓRICO

Desde o advento dos satélites observadores da Terra há várias décadas, diversos pesquisadores têm feito esforços para obter informações úteis para o setor agrícola usando essa fonte de informação. Um dos usos mais notórios dessa tecnologia na agricultura é para estimativas ou previsões de produtividade (ou seja, a produção de grãos – ou outro componente vegetal com valor comercial – por unidade de área). O processo de previsão de produtividade geralmente integra séries temporais de estatísticas históricas de produtividade e indicadores de produtividade, que podem ser provenientes de sensoriamento remoto (e.g. índices de vegetação), de modelos biofísicos, de medições de campo, etc. Esses indicadores são usados para parametrizar modelos de previsão usando critérios estatísticos (GALLEGO; CARFAGNA; BARUTH, 2010).

Existe uma considerável variedade de abordagens para estimar/predizer a produtividade de culturas agrícolas com dados provenientes de sensoriamento remoto (GALLEGO; CARFAGNA; BARUTH, 2010; MOULIN; BONDEAU; DELECOLLE, 1998). Abordagens mais simples são baseadas em relações empíricas entre a produtividade e os indicadores de produtividade usando tanto modelos paramétricos, como as regressões lineares (uni- ou multivariadas), ou não-paramétricos usando técnicas como *random forest*, *support vector machine* ou redes neurais artificias. Estudos pretéritos demonstram que estimativas de produtividade baseada em relações empíricas podem explicar até 80% da variabilidade na produtividade de culturas como milho e trigo para as áreas onde esses modelos foram ajustados (SHANAHAN et al., 2001; TUCKER; HOLBEN; ELGIN, 1980; WIEGAND; RICHARDSON, 1990). Entretanto, modelos puramente empíricos normalmente apresentam baixo grau de generalização o que compromete sua capacidade de extrapolação para diferentes locais ou anos.

Uma segunda classe de modelos usados para estimativas de produtividade estão relacionados aos estudos de Monteith (1977), os quais demonstram que a produção total de biomassa vegetal por unidade de área é proporcional à radiação fotossinteticamente ativa (RFA) absorvida pelas plantas durante a estação de crescimento. De acordo com Bloom et al.

(1985), a razão da biomassa pela RFA absorvida pelas plantas, conhecida como eficiência do uso da radiação (EUR) é relativamente constante, uma vez que as plantas ajustam à área foliar total em função de fatores limitantes de crescimento como estresses causados por falta de nutrientes ou temperatura. Diversos estudos têm confirmado à validade dessa abordagem, embora se reconheça que variações na EUR ocorrem em decorrência de diferentes fatores, especialmente quando as plantas passam por estresses hídricos (STEINMETZ et al., 1990). Modelos para estimativa de produtividade baseados no conceito de EUR possuem pelo menos quatro componentes de acordo com a equação 1:

$$Produtividade = \left(\sum_{t=1}^{n} fRFA_t \, xRFA_t \right) xEURxIC \qquad (1)$$

onde $fRFA_t$ é a fração RFA absorvida pela dossel vegetal no tempo *t*, IC é o índice de colheita, e EUR representa a eficiência no uso da radiação.

Uma terceira abordagem é a combinação informações de sensoriamento remoto provenientes de imagens de satélites e modelos de crescimento cultura (i.e., modelos mecanísticos – baseados em processos). O uso dos modelos de crescimento de plantas baseados em processos oferece uma nova perspectiva para o desenvolvimento de modelos de predição de produtividade com uma maior capacidade de generalização, uma vez que é possível considerar as complexas interações entre genética, ambiente e manejo no processo de previsão de produtividade. Como demonstrado em Sibley et al. (2014), existem, pelo menos, duas alternativas para combinar essas duas fontes de informação para previsão de produtividade de culturas agrícolas. A primeira delas é usar modelos baseados em processos para estimar a produtividade das culturas, com os dados de sensoriamento remoto empregados para ajustar os dados de entrada ou os parâmetros inciais do modelo, sendo esse último aplicado individualmente para cada pixel da imagem de satélite (CLEVERS, 1997; DENTE et al., 2008; DORAISWAMY et al., 2005; LAUNAY; GUERIF, 2005). Na prática, essa abordagem é comumente aplicada através da simulação de crescimento e produtividade de culturas para múltiplas combinações de fatores como datas de semeadura, densidade de plantas, genótipos, capacidade de retenção de água no solo, etc. Os valores simulados de variáveis como, índice de área foliar ou fRFA são comparados com estimativas derivadas de imagens de satélite para essas mesmas variáveis. Os dados de entradas e parâmetros que resultem na correspondência mais próxima entre os valores simulados e os observados (derivados de imagens de satélite) ao longo da estação de crescimento, resultando em um

quadrado médio do erro mais baixo por exemplo, são selecionadas, e a produtividade das culturas associada a essa simulação é atribuído ao pixel especificado.

A segunda alternativa é usar os modelos baseado em processos para gerar pseudo-observações que serão usadas para treinar modelos empíricos sob uma grande variedade de condições climáticas, de solo e de manejo e acessar a produtividade das culturas através de relações empíricas (AZZARI; JAIN; LOBELL, 2017; JIN et al., 2017; JIN; AZZARI; LOBELL, 2017; LOBELL et al., 2015; SIBLEY et al., 2014). Assim como na abordagem atenrior, várias simulações são realizadas com o modelo mecanístico de crescimento de cultura para diferentes combinações de dados de entrada e parâmetros. Porém, ao invéz de comparar diretamente os dados simulados com as estimativas derivadas das imanges de satélite, os dados simulados são usados para ajustar modelos de regressão que relaciona a produtividade à preditores como, índices de vegetação e variáveis meteorológicas. Um exemplo dessa abordagem é apresentado por Lobell et al. (2015) onde simulações de índice de área foliar provenientes de modelos de crescimento de culturas são convertidas para unidades de GCVI (*Green Chlorophyll Index*), usando relações empíricas descitas na literatura (NGUY-ROBERTSON et al., 2012), e posteriormente são utilizadas em conjunto com os dados simulados de produtividade para ajustar relações empíricas entre essas duas variáveis. Os modelos empíricos ajustados usando pseudo-observações de GCVI e produtividade são então aplicados para todos os pixlels da imagem de satélite. De acordo com Clevers (1997) essa segunda abordagem, apesar de sua maior simplicidade normalmente apresenta resultados que supera a primeira.

1.3 PROPOSIÇÃO

Apesar dos importantes avanços nos diferentes campos do sensoriamento remoto aplicado à estimativa/previsão de produtividade citados anteriormente, relevantes questões ainda necessitam ser estudadas com maior detalhamento, e portanto fazem parte dos objetivos desse trabalho.

Previsões de produtividade associadas à escalas geográficas mais extensas, como por exemplo à nível municipal, estadual, ou até mesmo nacional, requerem, além do entendimento da relação entre a variável resposta (produtividade) e dos preditores (e.g. índices de vegetação gerados a partir de imagens de satélite), um extenso conhecimento relacionado à localização e distribuição espacial das áreas produtores sobre a região considerada. Esse conhecimento é

essencial, pois apenas informações relevantes, associadas às áreas de produção agrícola precisam ser coletadas das imagens de satélites, descartando toda informação proveniente de outros alvos como cidades, florestas, corpos de água, etc. Além disso, essa informação precisa ser levantada durante o decorrer da estação de crescimento das culturas, adicionando um grau extra de complexidade aos modelos de previsões de produtividade em nível regional. Essas camadas de informações contento a localização espacial das áreas cultivadas com uma determinada cultura (e.g. soja) normalmente não são disponíveis ao acesso público na maioria dos países (incluindo o Brasil), ou não são disponibilizadas durante a estação de crescimento da cultura, como é o caso dos Estados Unidos, onde o Serviço Nacional de Estatística na Agricultura (*National Agricultural Statistic Service* – NASS), torna essa informação disponível com um ano de atraso em relação ao ano corrente. Dessa forma, o primeiro objetivo desse estudo foi o desenvolvimento e a validação de modelos capazes de identificar áreas agrícolas produtoras da cultura de interesse e predizer a produtividade dessas culturas à nível regional (nível municipal para o estado do Rio Grande do Sul – Brasil e nível de condado para o *Corn Belt* – Estados Unidos), usando apenas dados de domínio público, para a cultura da soja no Brasil e para a cultura do milho nos Estados Unidos como meses de antecedência em relação à colheita.

Outra lacuna explorada nesse estudo está relacionado à inclusão das variáveis meteorológicas nos modelos de previsão de produtividade. Variáveis dessa natureza têm uma grande contribuição na variabilidade da produtividade dentro e entre anos agrícolas e possuem um grande potencial para melhorar o desempenho dos modelos de previsão de produtividade (JOHNSON, 2014). Apesar disso, poucos estudos têm explorado os impactos dessas variáveis na assertividade das previsões (SHAO et al., 2015). Precipitação, temperatura média, máxima e mínima do ar, são as variáveis mais comumente incluídas nos modelos de previsão de produtividade (JOHNSON, 2014; SHAO et al., 2015). Outra variável com potencial para ser incluída nos modelos de previsão de produtividade é o deficit de pressão de vapor (DPV). O DPV é o gradiente entre o interior de folha saturado de vapor de água e o ar mais seco no exterior (ORT; LONG, 2014) e é amplamente utilizado como uma medida da demanda hídrica atmosférica que depende da temperatura e umidade do ar. O DPV tem sido frequentemente relatado como uma das variáveis meteorológicas mais importantes, explicando anomalias históricas de produtividade da cultura do milho em todo o meio-oeste americano (LOBELL et al., 2014).

Um terceiro aspecto que ainda é explorado de maneira incipiente nos modelos de predição de produtividade é relacionado à utilização de algoritmos mais complexos capazes

de captar e descrever com mais precisão as relações entre a variável resposta e os preditores. Nos últimos anos algoritmos de aprendizagem de máquina ou *machine learning*, como *random forest*, *support vector machine* e principalmente as redes neurais têm substituído as regressões lineares multivariadas na estimativa de produtividade de culturas agrícolas. Porém, com o advento da computação em nuvem e o significativo incremento da capacidade de armazenamento e processamento de dados, a possibilidade de utilização de algoritmos ainda mais sofisticados também evoluiu. As redes neurais de aprendizagem profunda (*Deep Learning Neural Networks)* representam uma classe relativamente nova de algoritmos baseados em redes neurais compostas de múltiplas camadas de processamento capazes de aprender representações complexas de dados usando múltiplos níveis de abstração. Esses algoritmos têm potencial para superar a maioria dos algoritmos anteriormente citados porém requerem grande quantidade de dados para ser adequadamente ajustados. Assim, o terceiro objetivo desse estudo é verificar a adequabilidade do uso de redes neurais de aprendizagem profunda para previsões de produtividade da cultura da soja, comparando seus resultados com algoritmos comumente usados (e.g. r*andom forest*, *support vector machine* e regressões lineares multivariadas).

Assim como as previsões de produtividade em escala regional são importantes para aspectos relacionados a logística, comercialização e criação de políticas agrícolas, previsões e estimativas de produtividade à nível de área agrícola são interessantes do aspecto relacionado ao manejo dentro do escopo da agricultura de precisão. Porém, quando se altera a escala geográfica em que o modelo será aplicado, necessita-se alterar a escala da coleta de dados de produtividade, a fim de calibrar os modelos de predição. A geração e coleta de dados à nível de área agrícola é uma atividade onerosa, que demanda tempo e mão de obra qualificada. Os mapas de colheita oferecem uma oportunidade interessante de coleta de dados nessa escala geográfica, porém existem diversos fatores que podem comprometer a qualidade desse produto (SCHWALBERT et al., 2018). Além disso, modelos de predição de produtividade calibrados em nível de área agrícola normalmente apresentam baixa capacidade de generalização tanto espacial quanto temporal. Assim, o quarto objetivo desse estudo foi explorar a capacidade de modelos empíricos de previsão de produtividade calibrados à nível de área agrícola, serem extrapolados espacial e temporalmente. Mais especificadamente, verificar se modelos desenvolvidos usando dados provenientes de áreas de uma determinada região em um determinado ano agrícola podem ser aplicados para: i) áreas provenientes de outras regiões com dados de produtividade seguindo uma distribuição de frequência semelhante, ii) áreas provenientes de outras regiões com dados de produtividade seguindo

uma distribuição de frequência não-semelhante, e iii) áreas provenientes de uma mesma região mas de um ano agrícola diferente daquele usado para parametrização do modelo.

1.3 MATERIAIS E MÉTODOS

Esta subseção apresenta uma versão simplificada dos materiais e métodos dos três artigos que compõem essa tese. Maior detalhamento das técnicas e análises utilizadas é apresentado nas subseções destinadas a esse propósito inseridas nas seções 2, 3 e 4.

Com exceção dos 19 mapas de colheita usados no terceiro artigo apresentado nessa tese, apenas dados e softwares de acesso público foram utilizados na elaboração deste estudo.

1..3.1 Materiais e métodos do artigo 1

Para o artigo 1 (Mid-season county-level corn yield forecast for us Corn Belt integrating satellite imagery and weather variables), NDVI (*Normalized Vegetation Index*) e EVI (*Enhanced Vegetation Index*) foram derivadas de imagens do sensor MODIS (*Moderate-Resolution Imaging Spectroradiometer*) embarcado no satélite Terra (coleções MODIS/006/MOD09Q1 e MODIS/006/MOD13Q1). Dados climáticos (temperatura, precipitação e DPV) foram acessados através do PRISM (*Parameter elevation Regression on Independent Slopes Model*) e GRIDMET (*Gridded Surface Meteorological Dataset*), dois bancos de dados contendo informações de clima na forma de grid georreferenciado. Todas as variáveis supracitadas foram coletadas durante um período de 10 anos (2008 até 2017) iniciando no dia 1 maio até o dia 20 de agosto de cada ano.

As informações da distribuição das áreas produtores de milho no *Corn Belt* americano foram acessadas através da CDL (*Cropland Data Layer)*, um banco de dados na forma de grid georreferenciado desenvolvido pelo NASS anualmente. Essas informações associadas as informações supracitadas foram usadas para treinar um modelo de classificação de cultura usando o algortimo r*andom forest* com o objetivo de identificar as áreas de milho durante a estação de crescimento em que a previsão de produtividade seria realizada. O modelo de classificação foi ajustado usando o tipo de cultura como variável dependente (e.g. soja, milho ou sorgo) e NDVI, EVI, precipitação, temperatura e DPV como variáveis independentes. Todas as informações usadas até esta etapa foram acessadas usando a plataforma GEE.

Dados de produtividade à nível de condado dos anos considerados nesse estudo foram coletados do banco de dados do USDA/NASS e foram usados para treinar um modelo empírico de predição de produtividade usando EVI, NDVI, precipitação, temperatura e DPV como preditores.

O modelos de classificação e regressão foram avaliados usando uma validação cruzada pelo método *leave-one-out.* Dessa forma 10 modelos diferentes foram ajustados, sempre removendo o ano que seria usada para validação. Os modelos de regressão foram avaliados usando o erro médio absoluto, a raiz do quadrado médio do erro, e a eficiência de Nash-Sutcliffe. O modelo de classificação foi avaialdo usando a acurácia global. Por último, uma análise de sensibilidade foi realizada para verificar com que antecedência o modelo preditivo pode ser implementado e qual o impacto da antecipação no desempenho geral do modelo. Dessa maneira, as variáveis foram subsequentemente removidos dos modelos de classificação e regressão e a mesma abordagem de validação mencionada acima foi usada. Os modelos foram testados usando dados até 11 de julho, 19 de julho, 27 de julho, 4 de agosto, 12 de agosto e 20 de agosto.

1.3.2 Materiais e métodos do artigo 2

Para o artigo 2 (Satellite-based soybean yield forecast: integrating machine learning and weather data for improving crop yield prediction in southern Brazil), imagens do sensor MODIS embarcado no satélite Terra (coleções MODIS/006/MOD09Q1 e MODIS/006/MOD13Q1), foram uadas para calcular dois índices de vegetação: NDVI e EVI. Temperatura da superfície do dossel vegetal foi acessada através do sensor MODIS embarcado no satélite Aqua (coleção MYD11A2). Precipitação foi acessada do banco de dados CHIRPS (*Climate Hazards Group Infrared Precipitation with Stations*). Todas as variáveis supracitadas foram coletadas para um período de 14 anos (2003 até 2016) iniciando no dia 5 março até o dia 15 de outubro. Todas as informações usadas até esta etapa foram acessadas usando a plataforma online GEE.

As informações da distribuição das áreas destinadas à agricultura para o estado do Rio Grande do Sul foram acessadas através do banco de dados georreferenciado disponibilizado pelo Cadastro Ambiental Rural (CAR).

Dados de produtividade à nível municipal de anos anteriores foram coletados do banco de dados do Instituto Brasileiro de Geografia e Estatística – IBGE/Sistema IBGE de Recuperação Automática – SIDRA (https://sidra.ibge.gov.br/pesquisa/pam/tabelas) e foram

usados para treinar um modelo empírico de predição de produtividade usando EVI, NDVI, precipitação e temperatura da superfície do dossel como preditores. Três algoritmos foram testados: regressão linear multivariada, r*andom forest*, e redes neurais de aprendizagem profunda (*Deep Learning Neural Networks*).

Os três algoritmos foram avaliados usando uma validação cruzada pelo método *leave-one-out.* Dessa forma 14 modelos diferentes foram ajustados, sempre removendo o ano que seria usada para validação. A performance do modelo foi avaliada usando o erro médio absoluto e a raiz do quadrado médio do erro. Por último, uma análise de sensibilidade foi conduzida para verificar com que antecedência o modelo preditivo pode ser implementado e qual o impacto da antecipação no desempenho geral do modelo. Dessa maneira, as variáveis foram subsequentemente removidos dos modelos regressão e a mesma abordagem de validação mencionada acima foi usada para avaliar os modelos. Os modelos foram testados usando dados até 16 de janeiro, 1 de fevereiro, 17 de fevereiro e 5 de março.

1.3.3 Materiais e métodos do artigo 3

Produtividade da cultura do milho em nível de área agrícola foi acessada através de 19 mapas de colheita (6 para o estado do Rio Grande do Sul, 7 para o estado do Mato Grosso e 6 para o estado do Kansas – EUA) para os anos de 2016 e 2017. Imagens do satélite Sentinel 2 foram recuperadas para essas áreas de maneira a calcular os seguintes índices de vegetação: NDVI, NDRE (*Normalized Difference Red Edge Index*) e GNDVI (*Green Normalized Difference Vegetation Index*). A data de coleta das imagens variou de uma área para outra dependendo da localização espacial e da incidência de nuvens. Procurou-se coletar imagens em um período compreendido entre 20 dias antes e 20 dias após o florescimento. Uma provável data de florescimento foi estimada baseada na data de plantio e colheita das áreas.

Modelos de regressão linear multivariada foram ajustados considerando produtividade como variável dependente e os índices de vegetação como variáveis independentes. O banco de dados foi dividido em dados de treinamento e validação. Modelos foram inicialmente validados localmente (validados para o ano e local onde foram inicialmente ajustados) e depois foram validados temporalmente (modelos ajustados com dados de 2016 aplicados em 2017 – apenas para o estado do Kansas) e espacialmente (modelos ajustados no Rio Grande do Sul, aplicados para o Mato Grosso e para o Kansas).

2 ARTIGO 1 – MID-SEASON COUNTY-LEVEL CORN YIELD FORECAST FOR US CORN BELT INTEGRATING SATELLITE IMAGERY AND WEATHER VARIABLES

Abstract

Yield estimations are of great interest to support interventions from governmental policies and to increase global food security. This study presents a novel model to perform in-season corn yield predictions at the US county-level, providing robust results under different weather and yield levels. The objectives of this study were to: i) evaluate the performance of a random forest classification to identify corn fields using NDVI, EVI and weather variables (temperature, precipitation, and vapor pressure deficit- VPD), ii) evaluate the contribution of weather variables when forecasting corn yield by using remote sensing data and perform a sensitivity analysis to explore the model performance in different dates, and iii) develop a model pipeline for performing in-season corn yield predictions at county-scale. Main outcomes from this study were: i) high accuracy (87% on average) for corn field classification achieved in late August, ii) corn yield forecasts with a mean absolute error (MAE) of 0.89 Mg ha^{-1}, iii) weather variables (VPD and temperature) highly influenced the model performance, and iv) model performance decreased when predictions were performed early in the season (mid-July), with MAE increasing from 0.87 Mg ha^{-1} to 1.36 Mg ha^{-1} when forecast timing changed from DOY 232 to DOY 192. This research portrays the benefits of integrating statistical techniques and remote sensing to field survey data in order to perform more reliable in-season corn yield forecasts.

Keywords: Crop classification, random forest, satellite imagery, yield forecast, MODIS.

Abbreviations: CDL, Cropland Data Layer; DOY, day of year; EVI, Enhanced Vegetation Index; GEE, Google Earth Engine; MODIS, Moderate Resolution Imaging Spectroradiometer; NASA, National Aeronautics and Space Administration; NASS, National Agricultural Statistic Service; NDVI, Normalized Difference Vegetation Index; RMSE, Root Mean Square Error; USDA, United States Department of Agriculture; VI, vegetation index; VPD, Vapor Pressure Deficit.

INTRODUCTION

Yield forecasts with high accuracy before harvest are extremely useful in agricultural decision-making processes, but its applicability largely depends on the spatial scale in which predictions are performed. On the one hand, within-field yield variability predictions are helpful to understand how crops respond to numerous management and environmental factors (Peralta et al., 2016; Lobell, 2013). On the other hand, yield forecast models at a larger scale (e.g. county, state and country) are useful for questions involving global food security, government assistance in food policies, and trade of agricultural commodities. Moreover, such forecasts can permit grain traders to make informed decisions, especially in food exporting countries such as the US (Sakamoto et al., 2014).

Remotely sensed vegetation indices (VIs) such as the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI) are commonly used for agricultural mapping and yield forecasting (Maselli and Rembold, 2001; Mkhabela et al., 2005; Funk and Budde, 2009). Researches have been focused on a wide range of satellite imagery data to predict corn yield worldwide (Bognár et al., 2011; Lobell, 2013; Hamada et al., 2015; Peralta et al., 2016; Schwalbert et al., 2018), demonstrating the potential of the yield forecast models based on remote sensing data, as a tool for providing quantitative and timely information on agricultural crops. Satellite images with greater spatial resolution, such as the freely available Landsat 8 (30m) and Sentinel 2 (10m) or commercial options RapidEye (5m) and Skysat (2m), are needed for field-level crop monitoring and yield forecasts. On the other hand, models designed to perform predictions on county, state or country level are based on images with an intermediate to coarse resolution, such as AVHRR (1km) and MODIS (250m). These images have advantages in regard to their superior temporal revisit frequency and larger spatial coverage, which avoids problems with cloud interference (Rembold et al., 2013).

There are two equally important characteristics that should be considered in yield forecast models: i) accuracy of the forecasts and ii) timing when forecasts are performed. Usually, those two aspects are related, in order that predictions performed early in the season have lower accuracy (Hayes and Decker, 1996; Shanahan et al., 2001; Wall et al., 2008). Sakamoto et al. (2014) found high accuracy in US county- and state-level predictions using images from the beginning of the season, with the error decreasing as more images were added during the progression of the crop growing season. Early yield predictions are highly affected by weather events (e.g. heavy precipitation, drought and heat stresses) and corresponding agronomic management decisions in the remaining growing season. Weather has a large contribution to yield variability within- and between-season and usually yield forecast models have an improved performance when those variables are taken into account (Johnson, 2014). Despite of that, only a few studies have examined impacts of these additional input variables on the model performance (Shao et al., 2015). Precipitation, daily average, maximum and minimum air temperature (based on weather stations), and daytime and nighttime land surface temperature (derived from earth observations) are the most common variables included into crop yield forecast models (Johnson, 2014; Shao et al., 2015). Another potential variable that can be included into yield forecast models is the vapor pressure deficit (VPD). The VPD is the gradient between the water vapor–saturated leaf interior and the drier bulk air (Ort and Long, 2014) and is widely used as a measure of atmospheric water demand that depends on air temperature and humidity. It has frequently been reported as one of the

most decisive weather variables on historical corn yield anomalies across the US Corn Belt (Lobell et al., 2014).

Currently, one of the most critical steps to make in-season yield predictions at large scales is related to obtaining reliable information about geographical distribution of field and crop yields across large areas (Sakamoto et al., 2014; Jin et al., 2017a; Shelestov et al., 2017). For the US, the National Agriculture Statistical Service (NASS), a statistical arm of the United States Department of Agriculture (USDA), releases a layer with detailed information about geographical distribution of fields since 1997 (since 2008 for the entire country). This information is usually released three months after the harvest of summer crops in the US (approximately early February). This information is valuable for training yield models but not useful for in-season ("near real-time") yield forecasts. Recent agricultural studies using remote sensing data have focused on exploring techniques aiming crop classification based on satellite images (Sakamoto et al., 2014; Jin et al., 2017a; Shelestov et al., 2017). This is an essential step on the development of near real-time forecast models. Classification trees techniques such as random forest is growing in popularity among the classification methods to crop mapping, presenting a high computational efficiency and robustness against overfitting (Belgiu and Drãgut, 2016). For this technique, users need to set only two parameters – the number of variables in the random subset at each node and the number of trees in the forest – and the output usually is not very sensitive to their values, avoiding any subjectivity (Liaw and Wiener, 2002).

The objectives of this study were to: i) evaluate the performance of a random forest classification to identify pixels where corn is grown at 250m resolution using NDVI and EVI derived from MODIS images and weather variables, ii) assess the impact of weather variables (precipitation, temperature, and VPD) on corn yield predictions and evaluate the model performance when forecasting corn yield earlier on the season, and iii) develop a model pipeline to perform corn yield forecast at county level for the US Corn Belt (e.g., Kansas, Iowa, and Indiana). For our analyses, we select Iowa, Indiana, and Kansas – the 1st, 4th, and 12th ranked US states for state-level average corn yield (2008-2017 average), respectively – to test the model at varying yield levels.

MATERIALS AND METHODS

Data sources

Historical county-level corn yield data (2008-2017) was obtained from the USDA/NASS ("https://quickstats.nass.usda.gov/"). This database is released as point

information in a county (each point is a county/year yield record) without geographical identification such as latitude and longitude.

Additionally, VIs from satellite imagery were obtained from MODIS Surface Reflectance products via the Google Earth Engine (GEE) platform (Gorelick et al., 2017). Since we are working on a large scale, and we need to build a crop land layer free of clouds for the entire region, the available options for satellite data were limited. The NASA Earth Observing System Data and Information System (EOSDIS) provided 8- and 16-days imageries on a near real-time basis allowing to retrieve satellite data with a minimal interference of clouds. This cloud-freeness is the main reason to choose the EOSDIS data for building our model. From those layers we retrieved two VIs, NDVI and EVI. The NDVI is a widely used VI, with several applications in agriculture including crop classification in the US Corn Belt (Wardlow and Egbert, 2008), however its ability to separate corn from soybean has been questioned since those crops have relatively similar NDVI profiles (Shao et al., 2010; Gonzalez-Sanchez et al., 2014) and within-crop variations of season are at least as large as inter-crop differences. Moreover, when corn and soybeans reach their peak growth stage and thus high biomass, NDVI usually saturates and no further allows for deciphering of differences in biomass. For that reason, we have included the EVI, since it is more sensitive in capturing variability during high-biomass periods (Zhong et al., 2016), despite of its lower image frequency (16 days) compare to the NDVI layers (8 days).

All NDVI images were generated using data from the collection MODIS/006/MOD09Q1. This collection provides images with 250-meter resolution, and each MOD09Q1 pixel contains the best possible observation during an 8-day period in order to minimize problems with cloud interference. All EVI images were obtained from the collection MODIS/006/MOD13Q1 that provides images with 250-meter resolution, and each MOD13Q1 pixel contains the best possible observation during a 16-day period. All the images from these two collections were gathered between May 1 and August 20 from 2008 to 2017. The starting date was selected based on the corn planting date. The initial date was defined in order to capture information from the beginning of the crop growing season in the Corn Belt, which is typically from May to October (USDA/NASS), up to the date of the yield forecast. The initial date is similar to the one used by Johnson et al. (2014) for the US Corn Belt. Moreover, Shanahan et al. (2001) shows that satellite images earlier than May 1st have a weak correlation with the final yield. The final date, August 20, was chosen in order to get images which cover the period of the highest reflectance of the corn canopy and where the selected weather variables have the highest correlation with the corn yield. This period is expected to

capture the two most important phenological stages, flowering and grain filling, which is usually in July and August (Johnson, 2014; Lobell 2015; Peng et al., 2018). Despite the use of fixed dates for forecasting yield in a large region as the US Corn Belt, this simple approach has produced robust results for other related studies (Schlenker and Roberts, 2009; Bolton and Friedl, 2013; Johnson, 2014; Sakamoto et al., 2014, Peng et al., 2018).” Because of the high variability related to planting date, comparative relative maturity (CRM), and management, fixed dates seems to be the more robust approach.

The Cropland Data Layer (CDL) was used in this study to retrieve information related to corn and non-corn field locations. The CDL is a raster, geo-referenced, 30-meter resolution, crop-specific land cover data layer created annually for the US using moderate resolution satellite imagery (e.g. Landsat and MODIS) and extensive on-the-ground agricultural measurements. Its accuracy exceeds 90% for crops such as corn and soybean (Johnson and Mueller, 2010). All CDL images between 2008 and 2017 were obtained from GEE platform. For this study, the CDL was re-projected to the MODIS sinusoidal projection and up-scaled to 250 m, so that all the pixels from CDL and from MODIS match perfectly. When changing the scale from 30 m to 250 m, the values of the new pixels were equal to the average from all the smaller pixels partly or entirely overlapping with the new pixel. This process was performed to identify the pure corn pixels.

Three weather variables were selected to be potentially included on the models: daily average temperature, precipitation and VPD. The first two are refer to negative correlations of heat and positive correlations of precipitation on corn yields (Smith, 1914; Wallace, 1920; Bolton and Friedl, 2013; Johnson, 2014). Additionally, the VPD is known for having strong influence in several processes during the crop growth (Messina et al., 2015; Basso and Ritchie, 2018) and can provide important information with potential to improve the model performance in years with large yield anomalies due to weather stress events. All the weather variables were summarized (averaged for temperature and VPD and summed for precipitation) on an 8 days period in order to exactly match with the NDVI derived from MODIS.

Temperature and precipitation were obtained from the Parameter-elevation Regressions on Independent Slopes Model (PRISM), and VPD was obtained from GRIDMET. Both layers are daily gridded datasets for the conterminous US, and provide information with a resolution of ~4 km. Thus those layers were re-projected and down-scaled in order to be combined with the rest of the collected information.

Data collection and organization

Before collecting data from the aforementioned sources, a mask layer was built which contains all the pixels with a high likelihood of overlapping corn fields in any growing season (pixels entirely contained within corn fields). This mask layer was basically a mosaic of all the re-projected CDLs (process described above) from 2008 to 2017. The function of this mask layer was to reduce the number of non-agricultural pixels in the crop classification step. The inclusion of this layer has significantly decreased the processing time. In addition, the layer increase the model accuracy due to lower variability in the input data. For each year, all pixels that were labeled as corn, at least once in a period between 2008 and 2017, were considered as candidates to overlap corn fields. All the collected information comprises the first step on the model development (Figure 1 – Step 1).

Crop classification

The second step on the model development was to train a model capable to differentiate corn from non-corn field pixels (Figure 1 – Step 2). This step was necessary because otherwise the model would become largely dependent on the CDL updates, commonly released three months after the harvest of US summer crops (early February), impeding any near real-time corn yield forecast. The crop classification model was based on the random forest algorithm. Random forest is an ensemble classifier that randomly selects a subset of training samples and variables to produce multiple decision trees. A larger fraction of the entire dataset (usually around two-thirds of the samples) is used to train the trees and the remaining fraction is used in a cross-validation technique for estimating how well the resulting random forest model performs (Breiman, 2001). This technique has become common in the remote sensing community due to the accuracy of its outcomes (Belgiu and Drãgut, 2016). The algorithm was set to use 600 trees, with a minimum leaf sample size of five to build the classification tree through the randomForest package (Liaw and Wiener, 2002) by the R program (R Core Team, 2017)

The classification model used crop types as the dependent variable (only two classes were considered, corn – 100% pure corn pixels – and non-corn). Factors such as NDVI, EVI, VPD, temperature, and precipitation were all considered as independent variables. All independent variables were only used up to the forecasting time within the season, therefore different classification models were trained for the different yield forecast dates. This implies that the crop mask can slightly change between different forecasting days. The classification model was run eight times following an assembly approach. In the first round only the multi-

temporal VIs were used. In the rounds number 2, 3, and 4, the weather variables were included individually into the model, then in rounds 5, 6, and 7 two weather variables were included in pairs, and finally a full model using all the variables were tested. The efficiency of this step was obtained using a leave-one-year-out cross-validation (removing one year per round from the model and then using that year as the validation) and calculating the overall accuracy for each validated year. Overall accuracy was computed by dividing the number of correctly classified observations by the total number of observations derived from the confusion matrix. The best model was considered the model with the highest and constant accuracy over the ten years, and it was selected to be used on the step 3.

Empirical relationships between yield, vegetation indices and weather

For building the forecast model in step 3, only the pixels tagged as corn were used and these corn pixels were averaged to county level in order to be combined with the yield information from USDA/NASS. A multivariate model was fitted using corn yield as the independent variable. The dependent variables were added following the same assembly approach used for the classification model (Figure 1 –step 3). This process was independently repeated for all the yield forecast dates considered in this study.

Model performance was evaluated using a leave-one-year-out cross-validation approach and four metrics were used to assess the model accuracy: the mean absolute error (MAE), the root-mean square error (RMSE), the bias coefficient and the Nash–Sutcliffe model efficiency coefficient (NSE). The MAE represents the average magnitude of the errors while RMSE is a quadratic scoring rule for the average magnitude of the error, and it is more useful when large errors are particularly undesirable. The RMSE will always be larger or equal to the MAE; the greater difference between them, the greater the variance in the individual errors in the sample. If the RMSE = MAE, then all the errors are of the same magnitude. Bias computes the average amount by which observed is greater than predicted, if the model is unbiased the index should be close to zero, positive values means that the model is underestimating the observed data and negative values means that observed values are overestimated. The NSE is a normalized statistic that determines the relative magnitude of the residual variance compared to the measured data variance and shows how well the prediction fits to the year-to-year yield variability, and its interpretation is analogous to the coefficient of determination (R^2).

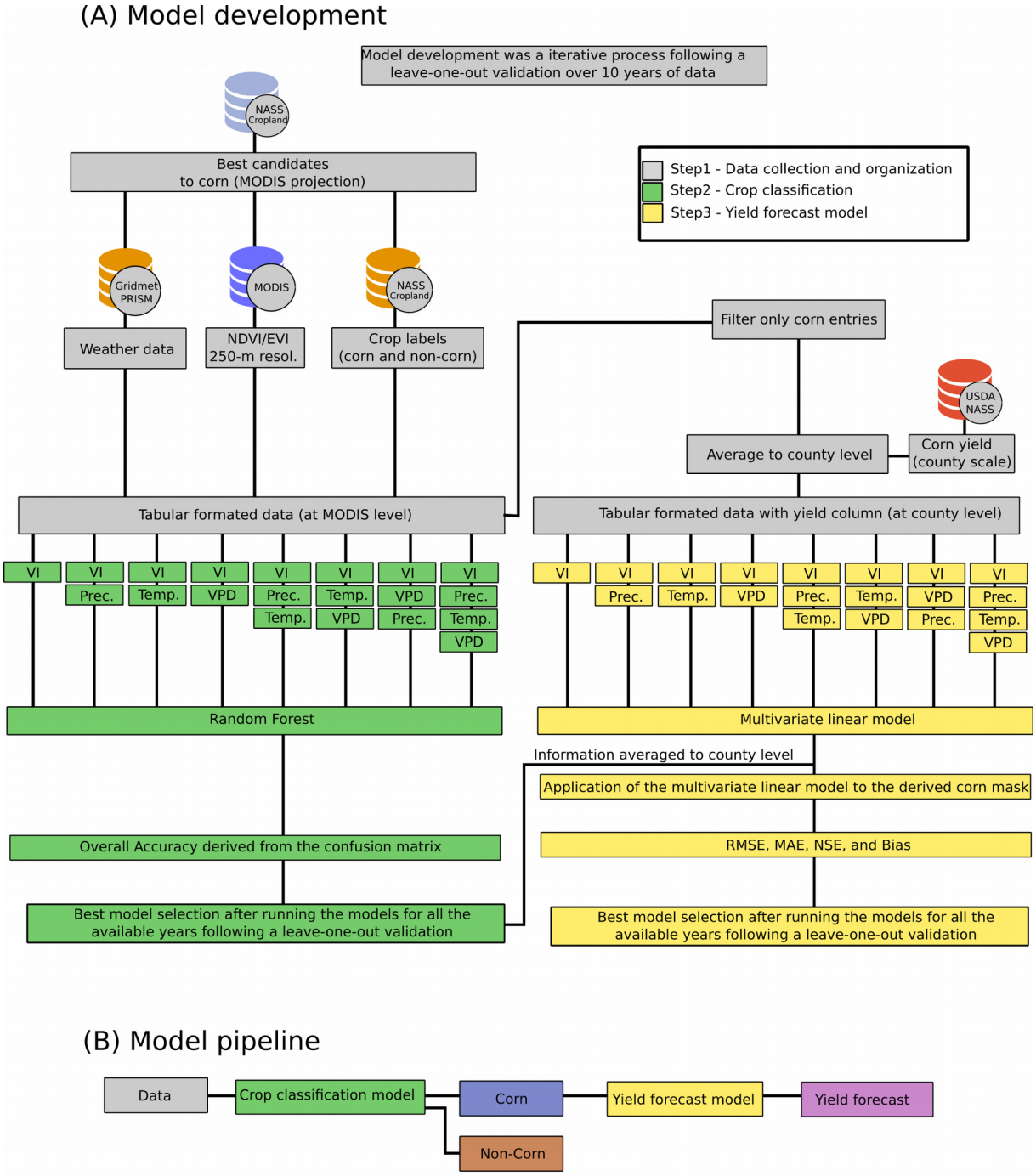


Figure 1. Flowchart indicating all steps of the (A) model development: 1 – data collection and organization, 2 – variable selection and crop classification model aiming at separating corn fields from non-corn fields using random forest algorithm (crop labels is referred to the labels used in the supervised classifications – random forest), step 3 – data selection and yield forecast model based on empirical relationships; and (B) pipeline with all steps for applying the model in future conditions.

Time series sensitivity analysis

After selecting the best models for crop classification and yield forecast, a sensitivity analysis was performed to check how early in the season the forecasting yield model can be implemented and its impact on the overall model performance. For this purpose, data collected later during the growing season were subsequently removed from the model and the same validation approach aforementioned was used to compute using MAE, RMSE, bias, and NSE. Thus, we tested the model using data until DOY 232 (August 20), DOY 224 (August 12), DOY 216 (August 4), DOY 208 (July 27), DOY 200 (July 19), and DOY 192 (July 11). We have assumed the existence of a delay in the release of the yield forecast models based on the process for uploading the MODIS product by NASA, 5 days (Sakamoto et al., 2014), and a processing time for the proposed algorithm under parallel processing to be 1 day, totalizing a delay of 6 days.

All the data collection and organization was performed on the GEE platform. The analysis comprised in step 2 and 3 were performed in the R environment in the Beocat, the High Performance Computer from Kansas State University, under parallel processing.

RESULTS

Crop classification

The importance of the weather on the crop classification was assessed using an assembly approach where the weather variables were independently added to the model. There was no significant improvement on the model performance after the inclusion of precipitation, temperature or VPD, compare to the model using only NDVI and EVI. Accuracy of the model presented some degree of variability over the growing seasons and among the states, with Iowa and Kansas having higher accuracy relative to Indiana (Figure 2). Additionally, 2012 had the lowest accuracy related to the other years considered on this study (lowest points in Figure 2A). In overall, average for all the years on the three states, the model presented an accuracy of 87% for DOY 232 (Figure 2), 87% for DOY 224, 86.5% for DOY 216, 86% for DOY 208, 84.6% for DOY 200, and 82% for DOY 192 (data not shown) related to the CDL.

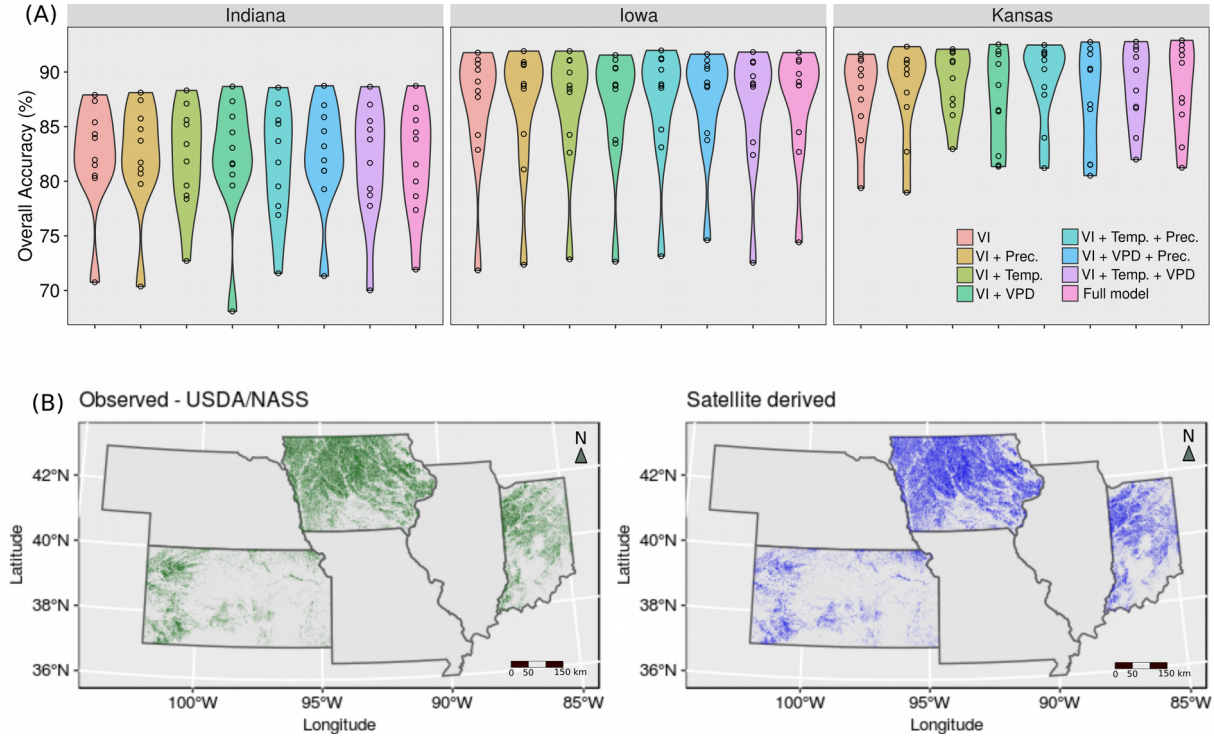


Figure 2. (A) Overall out-of-sample accuracy (number of correctly classified observations divided by the total number of observations derived from the confusion matrix), for all the years and states considered in this study (each point represents a year/state). The shapes around the points represent the kernel density plot for the accuracy in each condition. (B) Thematic maps representing the spatial distribution of corn pixels from CDL (at MODIS scale) (green) and predicted by the model using NDVI and EVI (blue) for 2017.

Empirical relationships between yield, vegetation indices and weather

There was a significant difference on the model performance driven by the inclusion of weather variables into the model. The simplest model, based only on NDVI and EVI resulted in the highest MAE (1.33 Mg ha^{-1}), RMSE (1.04 Mg ha^{-1}), the lowest NSE (0.7), and the most negative bias (-77 kg ha^{-1}), indicating that this model presented a large dispersion of points along the 1:1 line, and tended to overestimate the observed yield in a higher proportion compare to the other models. Following a hierarchical order of importance, the weather variable with the highest contribution on enhancing the model performance was VPD, followed by temperature and then accumulated precipitation. When the weather variables were included in pairs, the combination of precipitation + temperature did not yield better results than the model that only included VPD as the weather variable. The remaining three models had a better performance relative to the previous ones, with the model containing NDVI, EVI, temperature, and VPD having a similar performance to the full model and

slightly better than the model including precipitation instead of temperature (Figure 3A). For that reason we selected the model including temperature and VPD as the weather variables, additional to the VI predictors, for being used in the last step. The inclusion of temperature and VPD resulted in a decrease of 0.15 Mg ha^{-1} in MAE, 0.18 Mg ha^{-1} in RMSE, 68 kg ha^{-1} in bias, and an increase of 0.07 in the NSE, related to the model that only included NDVI and EVI (Figure 3B). The model performance was quite stable over the years. The 2012 growing season presented the lowest model performance, mainly related to the dry conditions during this growing season (2012 had the lowest yield average among all the years considered in this study). On the other hand, 2009, 2011, and 2016 presented the better model performance according to the RMSE, MSE, NSE, and Bias metrics (supplementary figure 1).

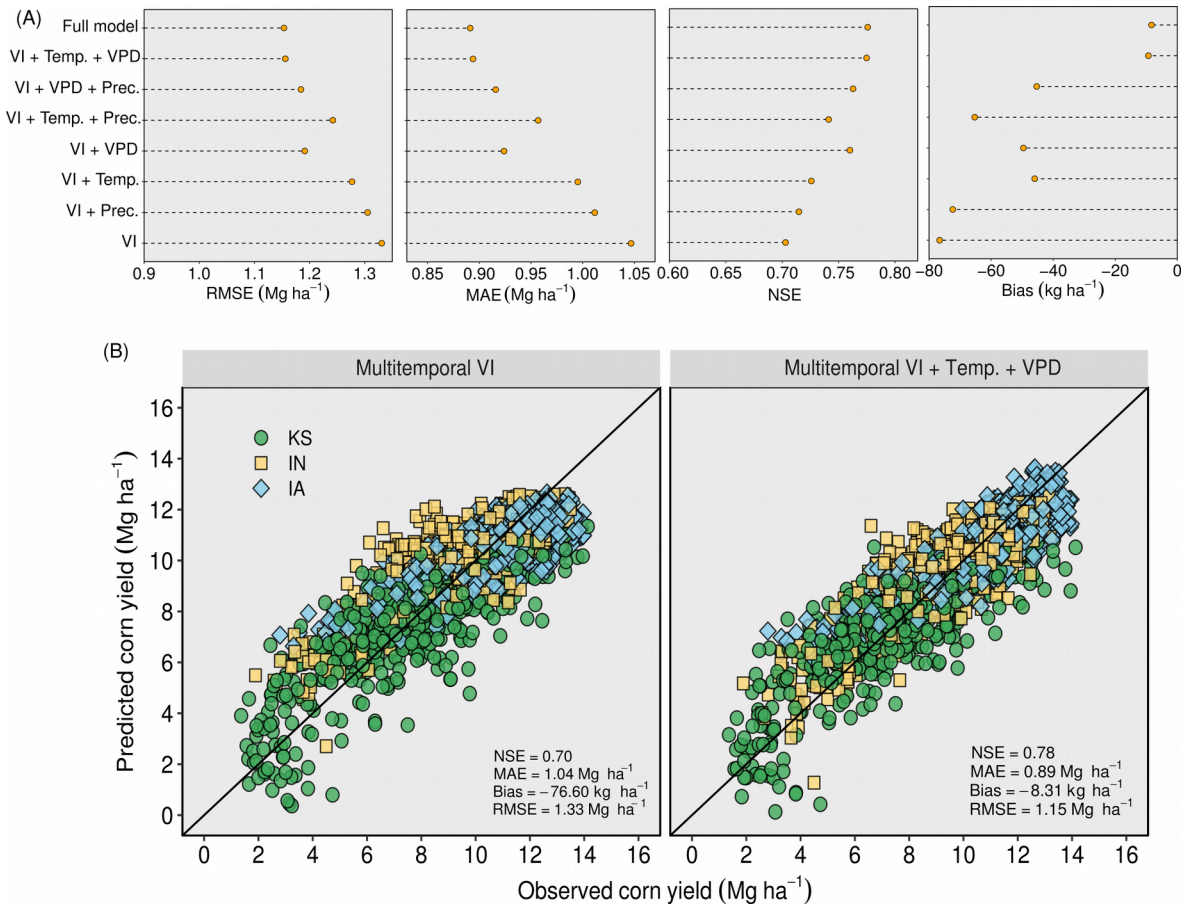


Figure 3. (A) Mean absolute error (MAE), Root-mean square error (RSME), Nash–Sutcliffe model efficiency coefficient (NSE), and bias coefficient for all the models tested in the assemble approach. (B) Observed versus out-of-sample forecasted corn yield from a yield forecast model with multitemporal VIs (left) and observed versus predicted corn yield from a yield forecast model with multitemporal VIs, temperature and vapor pressure deficit (right). Predictive yield based on aggregating data until DOY 232 (August 20) for Kansas, Indiana,

and Iowa from 2008 to 2017. The black line is presented in panel portraying the 1:1 line for the observed-predicted relationship. The sample size is n = 2501 data points.

Sensitivity of the results to forecasting time

The accuracy of the model decreased as the county-level corn yield forecast was anticipated from the DOY 232 (August 20) to DOY 196 (July 11) (Figure 4A). MAE and RMSE increased and NSE decreased as the yield forecast was performed earlier in the season. Model performance was most affected when the predictions were performed before DOY 208 (July 27), with MAE overpassing 1 Mg ha^{-1} (Figure 4B).

Another negative effect of performing yield forecast earlier in the season was the trend to overestimate yields in a higher frequency, evidenced by the decreased (more negative values) in the bias coefficient as the predictions are performed towards the beginning of the growing season. This behavior was more evident for the lowest yields on DOY 200 (July 19) and 196 (July 11).

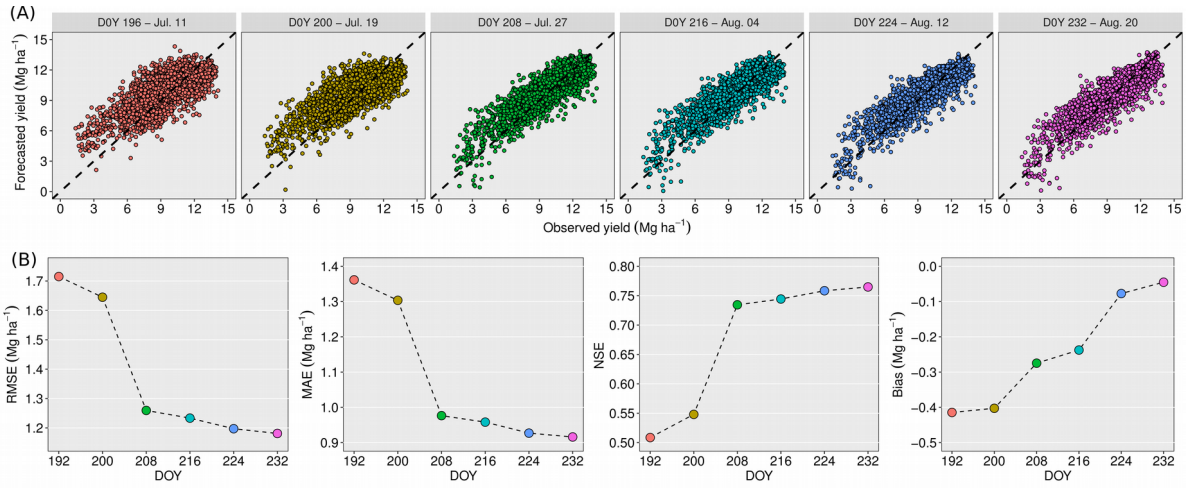


Figure 4. (A) Observed versus out-of-sample forecasted corn yield (forecast model with multitemporal VIs, temperature and vapor pressure deficit) for different dates expressed in days of year (DOY). A black dashed line is presented in panel portraying the 1:1 line for the observed-predicted relationship. The sample size is n = 2501 data points. (B) Variations in the mean absolute error (MAE), root-mean square error (RMSE), Nash–Sutcliffe model efficiency coefficient (NSE), and bias coefficient for different dates of yield prediction.

DISCUSSION

This study offered a novel approach using the CDL as a ground truth layer for crop classification and remote sensing combined with weather data as predictors for estimating corn yield at the county-scale. Moreover, it provides information related to the impact on model accuracy by anticipating corn yield forecast earlier in the season relative to the projected by USDA/NASS. This study suggests that at DOY 208 (July 27) models aiming at forecasting corn yield in the US Corn Belt could be implemented with an error (MAE) lower than 1 Mg ha^{-1}.

One of the main challenges for building accurate yield forecast models is to determine the geographical distribution of the fields, herein after termed as crop mapping layer. This information is valuable since all the pixels not corresponding to corn fields should be removed, “masked”, from the image, before establishing empirical relationships between VI and yield. Different techniques focused on crop classification and crop masking have been proposed worldwide, exploring differences in emergence dates between corn and soybean (Sakamoto et al., 2014), using different machine learning algorithms (e.g. supported vector machine, decision trees, and neural networks) (Shelestov et al., 2017), and exploring different resolution satellites options, such as Landsat 8, Sentinel 2, and RapidEye (Azzari et al., 2017; Jin et al., 2017a; Xiong et al., 2017). In the US the CDL is also an interesting option for masking crops pixels from satellite scenes, due its very high accuracy, exceeding 90% for corn and soybean (Johnson and Mueller, 2010). However this layer is not useful for near real-time yield forecast, since it is released with one year of delay. Despite of that, the CDL has great value as source of labeled data for training crop classification models. As the first outcome this study presented a novel crop classification approach based on Random Forest and multitemporal NDVI and EVI images from early May (DOY 128) to late August (DOY 232) and CDL from previous years as ground-truth for field geographical positions. This model achieved an accuracy of 87% (DOY 232) on average, higher than the 85% threshold, desired for most of agricultural purposes (Shelestov et al., 2017). Similar values of accuracy were reported in previous studies using MODIS images and different approaches for classifying the fields (Sakamoto et al., 2014). The crop classification model makes the yield forecast not dependent on the CDL updates and allows near real-time yield predictions.

The second outcome from this research was the development of a model to forecast corn yield at DOY 232 at county-level. The USDA/NASS usually releases the first corn yield report approximately at the 224th DOY, but in a state-level, county-level yield information is only released in the following year (usually three months after harvest). Satellite imagery data are known to be a useful and a reliable information to forecast yield before harvest time

(Bolton and Friedl, 2013; Sakamoto et al., 2014; Johnson, 2014; Peralta et al., 2016; Jin et al., 2017a; b), and the simplest approach to estimate crop yields by establishing empirical relationships between end-season yield observations and mid-season VIs calculated from multispectral images (Moriondo et al., 2007; Wall et al., 2008; Bognár et al., 2011; Minuzzi and Lopes, 2015; Shao et al., 2015; Hamada et al., 2015; Peralta et al., 2016; Bu et al., 2017). The model based on multitemporal VIs was able to predict yield with a MAE of 1.04 Mg ha^{-1} (DOY 232) over ~2,500 combinations of county-years and have its model performance significantly improved by the introduction of temperature and VPD as predictors, reaching a MAE of 0.89 Mg ha^{-1} (DOY 232). Information related to inclusion of weather variables on empirical yield forecast models are still scarce on the literature. Johnson (2014) found negative correlation between daytime surface temperature and corn yields, but lack of improvement on the model performance was documented by including nighttime surface temperature or precipitation. Additionally, Shao et al. (2015) did not find any benefits by including precipitation, average daily, maximum and minimum air temperature into the model. However, our study is one of the few studies evaluating the performance of VPD along with multitemporal VIs as predictors for estimating corn yield at county-scale. Lobell et al. (2014) reported VPD in the third month after sowing, which is typically July for a field sown in early May, as the most influencing variable among other 19 weather variables explaining historical yield variations across the US Corn Belt. VPD is a widely used measure of atmospheric water demand. It is closely related to crop evapotranspiration and consequently has major impacts on crop growth and yields. It has been documented that the photosynthetic rate declines when atmospheric VPD increases (Quick et al., 1992; Hirasawa and Hsiao, 1999; Fletcher et al., 2007). It is because plants under high VPD conditions reduce stomatal conductance, which effectively saves water in the plant, at the cost of reduced carbon assimilation (Lobell et al., 2013).

Lastly, a sensitivity analysis was pursued to explore how early reliable county-level corn yield predictions can be accomplished. The importance of a yield prediction could be considered as a balance between its accuracy and the timing when the prediction are performed, considering that usually there is a trade-off between the error and the date of the prediction (Bolton and Friedl, 2013; Sakamoto et al., 2014; Shao et al., 2015). Our results showed that corn yield can be forecasted at county-scale for the US Corn Belt at DOY 208 with a MAE < 1 Mg ha^{1}, and a RMSE of 1.26 Mg ha^{-1}. Equal RMSE was reported by Johnson et al. (2014) when forecasting yield at DOY 305 using the CDL as the crop mask. Furthermore, the RMSEs reported in this study are within the range of the values reported by

Shao et al. (2015), and bellow the ones reported by Sakamoto et al. (2014), ranging from ~1.68 Mg ha^{-1} to ~1.76 Mg ha^{-1} at DOY 215 when performing prediction independently from CDL for 2002 and 2012. Therefore, these results represent a great prospect for anticipating the yield forecast in approximately two weeks related to the first USDA/NASS yield report at state level.

Despite of the model pipeline presented in this study being dependent only on remote sensing and weather data to forecast corn yield at the county-level, the layers used to train the crop classification model and to establish the yield-VI/weather empirical relationships came from extensive field surveys performed by USDA/NASS or analogous agencies. Therefore, the contribution of the research is to show the potential benefits of integrating statistical techniques and remote sensing data to standard approaches (field survey) to perform more reliable in-season yield forecasts. Moreover, it is worth acknowledging that the model developed in this study presents limitations that can be overcome in future studies. The first constraint is related to the resolution, since the MODIS pixel size is 250 m. Thus, fields below that resolution are blended with other fields and may therefore be inaccurately treated in the analysis. The second constraint is related to the model dependence on field survey data, since this study was developed for the US, the crop classification model was trained using the CDL. For countries where this type of information is not yet available extensive field surveys will be required to achieve high accuracy in the crop classification step. The model performance could still be enhanced by i) adding phenology information during the crop classification process (Bolton and Friedl, 2013), ii) exploring new sources of information combining better spatial and temporal resolutions, such as Sentinel-2, RapidEye, and Skysat, iii) exploring new direct indicators of photosynthesis (such as solar-induced florescence) that will be available in a near future (Drusch et al., 2017), iv) adding management information into the model scope such as selection of crop varieties, fertilizer, plant density, comparative relative maturity (CRM), or irrigation, and v) combining remote sensing information and crop models [i.e. mechanistic (process-based) models] output to enhance predictability power and increase the spatio-temporal limits of predictability. As summarized in Sibley et al. (2013), there are at least two approaches for combining these two sources of information for forecasting crop yields. The first one is to use crop simulation models to forecast crop yields, with the remote sensing data employed to adjust inputs or parameters for the model on a pixel-by-pixel basis (Clevers, 1997; Doraiswamy et al., 2005; Doraiswamy et al., 2005; Launay and Guerif, 2005; Dente et al., 2008). The second approach is to use crop models for training empirical models under a larger variety of weather, soil and management conditions and access the crop yield

through the empirical coefficients (Sibley et al. 2013, Lobell 2015; Azzari et al. 2017; Jin et al., 2017a;b). Both approaches result in models less dependent on third-party data such as the USDA/NASS, and more robust against weather anomalies such as the 2012 growing season. Results from this study suggest that remote sensing and weather variables (temperature and VPD) are valuable data sources to perform accurate near real-time county-level corn yield predictions even early in the season (late July), having potential to enhance and help to anticipate yield predictions from official government departments such as the USDA/NASS. Despite only three states were considered in this study, Iowa, Indiana and Kansas, we tested the model for additional random combinations of counties (from different states) and years and the estimated error was within the range reported in the result section.

CONCLUSIONS

Multi-temporal satellite imagery combined with weather variables can provide useful information allowing the development of models able to forecast and monitoring corn yield at early season (after flowering) at county-scale. A decrease in accuracy is expected by anticipating the yield predictions, but this study suggests that corn yield forecast based on satellite imagery, temperature and VPD could be implemented at 208 DOY (July 27) with an accuracy of 78%. This is ~16 days before the first corn yield report of the USDA/NASS (at state level) and approximately 122 days before the harvest. Additionally, the novel crop classification model developed in this study using the Random Forest classification technique was adequate to separate pixels from MODIS images between corn and non-corn fields with an overall accuracy higher than 85%.

The training and validation approach used in this study with data from different states and years was adequate to test the model performance in different weather and yield conditions. Despite the analysis being developed for the US, the general approach described can potentially be applied to other regions around the globe if a reasonable amount of survey data is available for building a solid crop mapping data layer. This could contribute to support agricultural decisions in regard to managing and transferring risks within the crop production. This can help farmers to plan interventions and enable governments and traders to adjust trading schemes and thus, avoid yield failures and food shortages.

Acknowledgments: This study was supported by CAPES Foundation, Ministry of Education of Brazil, Brasilia - DF, Zip Code 70.040-020, Aquarius project (http://w3.ufsm.br/projetoaquarius), and Kansas Corn Commission. This is contribution no.

18-789 from the Kansas Agricultural Experiment Station and process 88887.130848/2016-00 from CAPES.

Author Contributions: Rai A. Schwalbert led the statistical analysis and evaluation of the crop classification and forecasting yield model and wrote the paper; Telmo J. C. Amado, Luciana Nieto, Geomar Corassa, and Charles Rice contributed to the data discussion; Nahuel Peralta, Bernhard Schauberger, and Christoph Gornott contributed to the data analysis/discussion and writing of the paper; Ignacio Ciampitti led the study and contributed to the data analysis/discussion and writing of the paper.

Conflict of interest: The authors declare no conflict of interest.

REFERENCES

Alganci, U., M. Ozdogan, E. Sertel, and C. Ormeci. 2014. Estimating maize and cotton yield in southeastern Turkey with integrated use of satellite images, meteorological data and digital photographs. F. Crop. Res. 157: 8–19. doi: 10.1016/j.fcr.2013.12.006.

Azzari, G., Jain, M., Lobell, D.B., 2017. Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries. Remote Sens. Environ. 202, 129–141. https://doi.org/10.1016/j.rse.2017.04.014

Ban, H.Y., K.S. Kim, N.W. Park, and B.W. Lee. 2017. Using MODIS data to predict regional corn yields. Remote Sens. 9(1): 1–19. doi: 10.3390/rs9010016.

Basso, B., and J.T. Ritchie. 2018. Evapotranspiration in high-yielding maize and under increased vapor pressure deficit in the US Midwest. Ael 3(1): 0. doi: 10.2134/ael2017.11.0039.

Belgiu, M., and L. Drãgut. 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS J. Photogramm. Remote Sens. 114: 24–31. doi: 10.1016/j.isprsjprs.2016.01.011.

Bognár, P., C. Ferencz, S. Pásztor, G. Molnár, G. Timár, et al. 2011. Yield forecasting for wheat and corn in Hungary by satellite remote sensing. Int. J. Remote Sens. 32(17): 4759–4767. doi: 10.1080/01431161.2010.493566.

Bolton, D.K., and M.A. Friedl. 2013. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. Agric. For. Meteorol. 173: 74–84. doi: 10.1016/j.agrformet.2013.01.007.

Breiman, L. 2001. Random Forests. Mach. Learn. 45: 5–32. https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf (accessed 16 April 2018).

Bu, H., L.K. Sharma, A. Denton, and D.W. Franzen. 2017. Comparison of satellite imagery and ground-based active optical sensors as yield predictors in sugar beet, spring wheat, corn, and sunflower. Agron. J. 109(1): 299–308. doi: 10.2134/agronj2016.03.0150.

Dente, L., Satalino, G., Mattia, F., Rinaldi, M., 2008. Assimilation of leaf area index derived from ASAR and MERIS data into CERES-Wheat model to map wheat yield. Remote Sens. Environ. 112, 1395–1407. https://doi.org/10.1016/j.rse.2007.05.023

Doraiswamy, P.C., Moulin, S., Cook, P.W., Stern, A., 2003. Crop Yield Assessment from Remote Sensing. Photogramm. Eng. Remote Sensing 69, 665–674. https://doi.org/10.14358/PERS.69.6.665

Doraiswamy, P.C., Sinclair, T.R., Hollinger, S., Akhmedov, B., Stern, A., Prueger, J., 2005. Application of MODIS derived parameters for regional crop yield assessment. Remote Sens. Environ. 97, 192–202. https://doi.org/10.1016/j.rse.2005.03.015

Drusch, M., J. Moreno, U. Del Bello, R. Franco, Y. Goulas, et al. 2017. The FLuorescence EXplorer Mission Concept—ESA's Earth Explorer 8. IEEE Trans. Geosci. Remote Sens. 55(3): 1273–1284.

Fletcher, A. L., T. R. Sinclair, and L. H. Allen Jr., 2007: Transpiration responses to vapor pressure deficit in well watered "slow-wilting" and commercial soybean. Environ. Exp. Bot. 61, 145–151, doi: 10.1016/j.envexpbot.2007.05.004.

Funk, C., and M.E. Budde. 2009. Phenologically-tuned MODIS NDVI-based production anomaly estimates for Zimbabwe. Remote Sens. Environ. 113(1): 115–125. doi: 10.1016/j.rse.2008.08.015.

Gonzalez-Sanchez, A., J. Frausto-Solis, and W. Ojeda-Bustamante. 2014. Attribute Selection Impact on Linear and Nonlinear Regression Models for Crop Yield Prediction. Sci. World J. 2014(Ml): 1–10. doi: 10.1155/2014/509429.

Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, et al. 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202: 18–27. doi: 10.1016/j.rse.2017.06.031.

Hamada, Y., H. Ssegane, and M.C. Negri. 2015. Mapping intra-field yield variation using high resolution satellite imagery to integrate bioenergy and environmental stewardship in an agricultural watershed. Remote Sens. 7(8): 9753–9768. doi: 10.3390/rs70809753.

Hamar, D., C. Ferencz, J. Lichtenberger, G. Tarcsai, and I. Ferencz-Arkos. 1996. Yield estimation for corn and wheat in the Hungarian Great Plain using Landsat MSS data. Int. J. Remote Sens. 17(9): 1689–1699. doi: 10.1080/01431169608948732.

Hayes, M.J., and W.L. Decker. 1996. Using NOAA AVHRR data to estimate maize production in the United States Corn Belt. Int. J. Remote Sens. 17(16): 3189–3200. doi: 10.1080/01431169608949138.

Hirasawa, T., and T. C. Hsiao, 1999: Some characteristics of re- duced leaf photosynthesis at midday in maize growing in the field. Field Crop. Res., 62, 53–62, doi: 10.1016/S0378-4290(99)00005-2.

Jin, Z., G. Azzari, M. Burke, S. Aston, and D. Lobell. 2017a. Mapping Smallholder Yield Heterogeneity at Multiple Scales in Eastern Africa. Remote Sens. 9(9): 931. doi: 10.3390/rs9090931.

Jin, Z., G. Azzari, and D.B. Lobell. 2017b. Improving the accuracy of satellite-based high-resolution yield estimation: A test of multiple scalable approaches. Agric. For. Meteorol. 247: 207–220. doi: 10.1016/j.agrformet.2017.08.001.

Johnson, D.M. 2014. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. Remote Sens. Environ. 141: 116–128. doi: 10.1016/j.rse.2013.10.027.

Johnson, D.M., and R. Mueller. 2010. The 2009 Cropland Data Layer. Photogramm. Eng. Remote Sens. 76(11): 1201–1205.

Schlenker, W. and M. J. Roberts. 2009. Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. Proceedings of the National Academy of Sciences, 106(37), 15594–15598. doi:10.1073/pnas.0906865106

Launay, M., Guerif, M., 2005. Assimilating remote sensing data into a crop model to improve predictive performance for spatial applications. Agric. Ecosyst. Environ. 111, 321–339. https://doi.org/10.1016/j.agee.2005.06.005

Liaw, A., and M. Wiener. 2002. Classification and Regression by randomForest. R News 2(3): 18–22. http://cran.r-project.org/doc/Rnews/.

Lobell, D.B., Hammer, G.L. McLean, G. Messina, C. Roberts, M.J., Schlenker, W., 2013. The critical role of extreme heat for maize production in the United States. Nat. Clim. Chang. 3, 497–501. https://doi.org/10.1038/nclimate1832

Lobell, D.B. 2013. The use of satellite data for crop yield gap analysis. F. Crop. Res. 143: 56–64. doi: 10.1016/j.fcr.2012.08.008.

Lobell, D.B., M.J. Roberts, W. Schlenker, N. Braun, B.B. Little, et al. 2014. Greater sensitivity to drought accompanies maize yield increase in the U.S. Midwest. Science (80-.). 344(6183): 516–519. doi: 10.1126/science.1251423.

Maselli, F., and F. Rembold. 2001. Analysis of GAC NDVI data for cropland identification and yield forecasting in mediterranean african countries. Photogramm. Eng. Remote Sens. 67(5): 593–602. https://pdfs.semanticscholar.org/1f7d/5eba601324ab23d94f9178ec36f0643761e1.pdf?_ga=2.6029107.1677474057.1506541835-1374488681.1506541835 (accessed 27 September 2017).

Messina, C.D., T.R. Sinclair, G.L. Hammer, D. Curan, J. Thompson, et al. 2015. Limited-transpiration trait may increase maize drought tolerance in the US corn belt. Agron. J. 107(6): 1978–1986. doi: 10.2134/agronj15.0016.

Minuzzi, R.B., and F.Z. Lopes. 2015. Desempenho agronômico do milho em diferentes cenários climáticos no Centro-Oeste do Brasil. Rev. Bras. Eng. Agrícola e Ambient. 19(8): 734–740. doi: 10.1590/1807-1929/agriambi.v19n8p734-740.

Mkhabela, M.S., M.S. Mkhabela, and N.N. Mashinini. 2005. Early maize yield forecasting in the four agro-ecological regions of Swaziland using NDVI data derived from NOAA's-AVHRR. Agric. For. Meteorol. 129(1–2): 1–9. doi: 10.1016/j.agrformet.2004.12.006.

Moriondo, M., F. Maselli, and M. Bindi. 2007. A simple model of regional wheat yield based on NDVI data. Eur. J. Agron. 26(3): 266–274. doi: 10.1016/j.eja.2006.10.007.

Ort, D.R., and S.P. Long. 2014. Limits on yields in the Corn Belt. Science (80-.). 344(6183): 484–485. doi: 10.1126/science.1253884.

Peng, B., Guan, K., Pan, M., Li, Y., 2018. Benefits of seasonal climate prediction and satellite data for forecasting U.S. maize yield. Geophys. Res. Lett. 45, 9662–9671. https://doi.org/10.1029/2018GL079291

Peralta, N., Y. Assefa, J. Du, C. Barden, and I. Ciampitti. 2016. Mid-season high-resolution satellite imagery for forecasting site-specific corn yield. Remote Sens. 8(10): 1–16. doi: 10.3390/rs8100848.

Quick, W., M. Chaves, R. Wendler, et al., 1992: The effect of wa- ter stress on photosynthetic carbon metabolism in four spe- cies grown under field conditions. Plant Cell Environ., 15, 25–35, doi: 10.1111/pce.1992.15.issue-1.

R Core Team. 2017. R: A Language and Environment for Statistical Computing. https://www.r-project.org/.

Rembold, F., C. Atzberger, I. Savin, and O. Rojas. 2013. Using low resolution satellite imagery for yield prediction and yield anomaly detection. Remote Sens. 5(4): 1704–1733. doi: 10.3390/rs5041704.

Sakamoto, T., A.A. Gitelson, and T.J. Arkebauer. 2014. Near real-time prediction of U.S. corn yields based on time-series MODIS data. Remote Sens. Environ. 147: 219–231. doi: 10.1016/j.rse.2014.03.008.

Schwalbert, R.A., T.J.C. Amado, L. Nieto, S. Varela, G.M. Corassa, et al. 2018. Forecasting maize yield at field scale based on high-resolution satellite imagery. Biosyst. Eng. 171: 179–192. doi: 10.1016/j.biosystemseng.2018.04.020.

Shanahan, J.F., J.S. Schepers, D.. D. Francis, G.E. Varvel, W.W. Wilhelm, et al. 2001. Use of Remote-Sensing Imagery to Estimate Corn Grain Yield. Agron. J. 93: 583–589. doi: 10.2134/agronj2001.933583x.

Shao, Y., Lunetta, R.S., Ediriwickrema, J. and Iiames, J. 2010. Mapping Cropland and Major Crop Types across the Great Lakes Basin using MODIS-NDVI Data. Photogramm. Eng. Remote Sensing 75(1): 73–84. doi: 10.14358/PERS.76.1.73.

Shao, Y., J.B. Campbell, G.N. Taff, and B. Zheng. 2015. An analysis of cropland mask choice and ancillary data for annual corn yield forecasting using MODIS data. Int. J. Appl. Earth Obs. Geoinf. 38: 78–87. doi: 10.1016/j.jag.2014.12.017.

Shelestov, A., M. Lavreniuk, N. Kussul, A. Novikov, and S. Skakun. 2017. Exploring Google Earth Engine Platform for Big Data Processing: Classification of Multi-Temporal Satellite Imagery for Crop Mapping. Front. Earth Sci. 5: 17. doi: 10.3389/feart.2017.00017.

Sibley, A.M., P. Grassini, N.E. Thomas, K.G. Cassman, and D.B. Lobell. 2014. Testing remote sensing approaches for assessing yield variability among maize fields. Agron. J. 106(1): 24–32. doi: 10.2134/agronj2013.0314.

Smith, J.W. 1914. The effect of weather upon the yield of corn. Mon. Weather Rev. 42(2): 78–92. doi: 10.1175/1520-0493(1914)42<78:TEOWUT>2.0.CO;2.

Wall, L., D. Larocque, and P.-M. Léger. 2008. The early explanatory power of NDVI in crop yield modelling. Int. J. Remote Sens. 29(8): 2211–2225. doi: 10.1080/01431160701395252.

Wallace, H.A. 1920. Mathematical inquiry into the effect ofweather on corn yield in the eight Corn Belt states. Mon. Weather Rev. 48: 439–446.

Wardlow, B.D., and S.L. Egbert. 2008. Large-area crop mapping using time-series MODIS 250 m NDVI data: An assessment for the U.S. Central Great Plains. Remote Sens. Environ. 112(3): 1096–1116. doi: 10.1016/j.rse.2007.07.019.

Xiong, J., P.S. Thenkabail, J.C. Tilton, M.K. Gumma, P. Teluguntla, et al. 2017. Nominal 30-m cropland extent map of continental Africa by integrating pixel-based and object-based algorithms using Sentinel-2 and Landsat-8 data on google earth engine. Remote Sens. 9(10): 1065. doi: 10.3390/rs9101065.

Zhang, S., Tao, F., Zhang, Z., 2017. Spatial and temporal changes in vapor pressure deficit and their impacts on crop yields in China during 1980–2008. J. Meteorol. Res. 31, 800–808. https://doi.org/10.1007/s13351-017-6137-z

Zhong, L., L. Yu, X. Li, L. Hu, and P. Gong. 2016. Rapid corn and soybean mapping in US Corn Belt and neighboring areas. Sci. Rep. 6(1): 36240. doi: 10.1038/srep36240.

3 ARTIGO 2 – SATELLITE-BASED SOYBEAN YIELD FORECAST: INTEGRATING MACHINE LEARNING AND WEATHER DATA FOR IMPROVING CROP YIELD PREDICTION IN SOUTHERN BRAZIL

Abstract

Soybean yield predictions in Brazil are of great interest for market behavior, to drive governmental policies and to increase global food security. In Brazil soybean yield data generally demand various revisions through the following months after harvest suggesting that there is space for improving the accuracy and the time of yield predictions. This study presents a novel model to perform in-season ("near real-time") soybean yield forecasts in southern Brazil using Long-Short Term Memory (LSTM), Neural Networks, satellite imagery and weather data. The objectives of this study were to: i) compare the performance of three different algorithms (multivariate OLS linear regression, random forest and LSTM neural networks) for forecasting soybean yield using NDVI, EVI, land surface temperature and precipitation as independent variables, and ii) evaluate how early (during the soybean growing season) this method is able to forecast yield with reasonable accuracy. Satellite and weather data were masked using a non-crop-specific layer with field boundaries obtained from the Rural Environment Registry that is mandatory for all farmers in Brazil. Main outcomes from this study were: i) soybean yield forecasts at municipality-scale with a mean absolute error (MAE) of 0.24 Mg ha^{-1} at DOY 64 (march 5) ii) a superior performance of the LSTM neural networks relative to the other algorithms for all the forecast dates except DOY 16 where multivariate OLS linear regression provided the best performance, and iii) model performance (e.g., MAE) for yield forecast decreased when predictions were performed earlier in the season, with MAE increasing from 0.24 Mg ha^{-1} to 0.42 Mg ha^{-1} (last values from OLS regression) when forecast timing changed from DOY 64 (March 5) to DOY 16 (January 6). This research portrays the benefits of integrating statistical techniques, remote sensing, weather to field survey data in order to perform more reliable in-season soybean yield forecasts.

Keywords: Yield forecast; Satellite imagery, deep learning, Long-Short Term Memory.

Introduction

Soybean [*Glycine max* (L.) Merrill] represents one of the world's most important sources of protein and oil, with four countries, US, Brazil, Argentina, and China, accounting for approximately 90% of the total global production (Embrapa, 2018; USDA, 2019). Brazil is currently the second largest soybean producer, only behind the US, contributing to ~34.7% of the global production. As a consequence, the soybean production from Brazil has a large impact on the global market, with seasonal fluctuations on production impacting the financial market.

In Brazil, there are two institutions responsible for providing data about the status of the crops, the National Supply Company (Conab) and the Brazilian Institute of Geography and Statistics (IBGE). Both Conab and IBGE are primarly based on field survies and they release annually yield forecasts (before harvest) on a state-level and estimations (after harvest) on a municipality-level (the last is released only by IBGE). Alternatively, with the advent of new cloud platforms such as Google Earth Engine (GEE) (Gorelick et al., 2017) providing an easier way to access large volumes of satellite and weather data, and dramatically increasing processing power through parallel computing resources, satellite imagery became an easy alternative for providing yield forecasts over larger domains in a near real-time basis. Research have repeatedly shown the potential of satellite imagery on providing quantitative data about yield worldwide (Ferencz et al., 2004; Hamada et al., 2015; Lobell, 2013; Peralta et al., 2016; Schwalbert et al., 2018), and improved model performance has been documented when weather data is effectively integrated on the estimations (Cai et al., 2018; Johnson, 2014; Lobell et al., 2015; Peng et al., 2018).

Along with the increase in computational processing power, more complex algorithms to data analysis also have become more popular when exploring larger and spatio-temporal datasets. Empirical relationships between soybean yield, canopy reflectance, and weather data usually present non-linearities (Johnson et al., 2016), and yield forecast models using a collection of those variables recorded over time are prone to over-fitting due to a high degree of autocorrelation. For those reasons, machine learning algorithms are able to more robustly deal with non-linearities against over-fitting. Those machine learning algorithms such as random forest and the neural networks have been successfully utilized to predict crop yield using remotely sensed vegetation indices (Alvarez, 2009; Cai et al., 2018; Johnson et al., 2016; Khaki and Wang, 2019; Li et al., 2013; Drummond et al., 2013; Shao et al., 2015). Random forest is an ensemble classifier that bootstraps training samples and variables to produce multiple decision trees performing predictions after aggregating the results from individual trees; this process is also known as bagging (Breiman, 2001). The neural networks consist of layers of highly interconnected processing units (neurons). The data moves throughout those layers across weighed connections, and each inner neuron is associated with an activation function, usually responsible for a non-linear transformation (Cai et al., 2018). A specific variation of the neural network, known as Long-Short Term Memory (LSTM) has been more recently noticed because of its large capacity to deal with sequential data (Cunha et al., 2018; You et al., 2017).

In addition to data processing, another challenge when performing yield forecast over large domains is to access to the crop geolocations. For some regions of the world such as the US, this information is easily available since it is yearly released by the National Agricultural Statistic Service (NASS) named Cropland Data Layer.. A 30-m resolution crop specific gridded layer (Johnson and Mueller, 2010) that is largely employed as a relevant layer in studies aiming at forecasting crop yield in the US (Johnson, 2014; Shao et al., 2015). In Brazil, such information is not yet available, despite the efforts of the governmental agencies. However, for most of the municipalities (similar to the county-level in US) in Brazil, it is possible to access the field boundaries of permanent agricultural fields from the Rural Environmental Registry (Cadastro Ambiental Rural - CAR) (http://www.car.gov.br). This layer despite not holding information related to crop types, provide an useful data source for removing most part of the noise from the satellite imagery, coming from areas that are not meaningful for agricultural purposes.

Thus, considering the importance of soybean in Brazil and its impact on the global economy, and the evident lack of reliable yield information in near real-time basis, the implementation of a near-real time yield forecast will provide a useful layer for agricultural purposes and policy applications. Therefore, the objectives of this research were to: i) compare the performance of three different algorithms (multivariate ordinary least square – OLS - linear regression, random forest and LSTM neural network) for forecasting soybean yield using vegetation indices such as NDVI, EVI, and weather data such as land surface temperature and precipitation as independent variables, and ii) evaluate how early (during the soybean growing season) this method is able to forecast yield with reasonable accuracy.

Material and Methods

Region selection

The study was conducted in the northern region of the Rio Grande do Sul (RS) state, Brazil. This region was chosen due to: i) the high area and frequency of soybean crop in the soybean-corn summer crop rotation (85% of the cropland is allocated to soybean), and ii) since its represents the largest contiguous cropland area in RS state (Figure 1A).

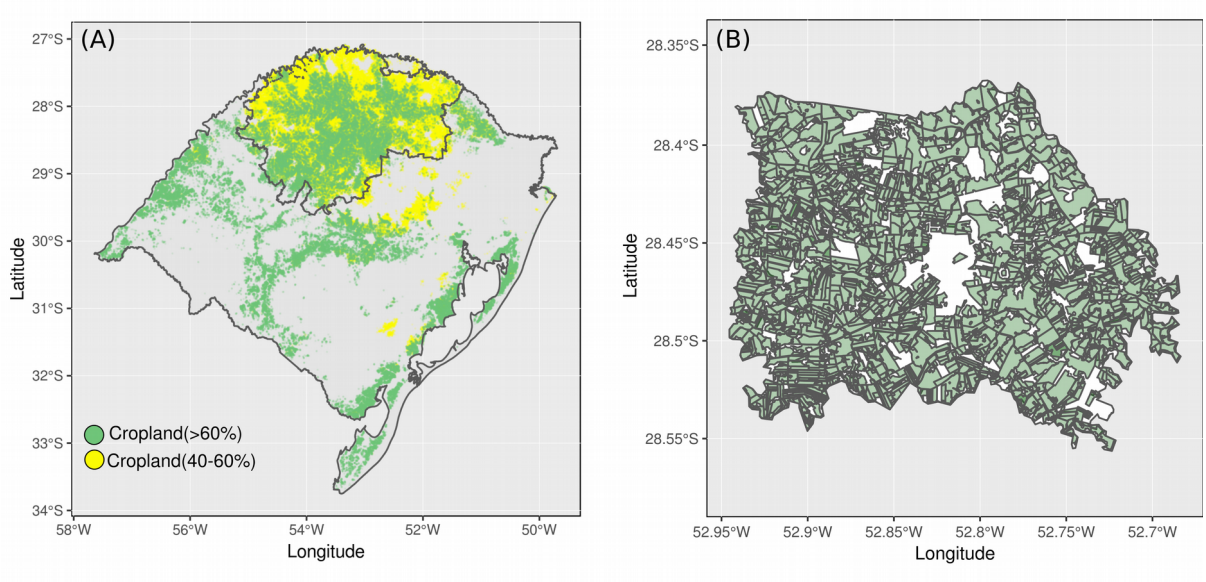


Figure 1. (A) Annual International Geosphere-Biosphere Programme (IGBP) land cover classification generated by NASA LP DAAC (500 m - spatial resolution). Only two classes (12 and 14 from original raster file) are highlighted, with percentage of the pixel covered with cropland ranging from 40 to 100%. (B) Example of file available for downloading in CAR - Consolidated areas for the municipality of Não-Me-Toque, RS.

Data sources

Historical municipality-level soybean yield data (2003-2016) was obtained from IBGE (https://sidra.ibge.gov.br/pesquisa/pam/tabelas). This database is released as a point information in a municipality (each point is a municipality/year yield record) without geographical identification such as latitude and longitude. We used 80 municipalities once we focused only in the ones with yield data available for the entire period considered in the study.

Additionally, vegetation indices (VIs) from satellite imagery were obtained from MODIS Surface Reflectance products. Since we are working on a large region, and we need to build a mosaic free of clouds for the entire region, the available options for satellite data are limited. The NASA Earth Observing System Data and Information System (EOSDIS) provided 8- and 16-days mosaics on a near real-time basis allowing to retrieve satellite data with minimal interference of clouds. This cloud-freeness is the main reason to choose the EOSDIS data for building our model. From those mosaics, we retrieved two VIs, the normalized difference vegetation index (NDVI), and enhanced vegetation index (EVI). Since EVI is released in a lower image frequency (every 16 days) compare to the NDVI (every 8

days) we calculated the average between each two consecutive EVI images in order to provide an EVI time series that matches with the NDVI images.

All NDVI images were generated using data from the collection MODIS/006/MOD09Q1. This collection provides images with 250-m resolution, and each MOD09Q1 pixel contains the best possible observation during an 8-day period in order to minimize problems with cloud interference. All EVI images were obtained from the collection MODIS/006/MOD13Q1 that provides images with 250-m resolution, and each MOD13Q1 pixel contains the best possible observation during a 16-day period. All the images from these two collections were gathered between October 15 and March 5 (soybean planting and harvesting are not in the same calendar year in Brazil) from 2002 to 2016. The starting date was selected based on the soybean planting date and phenology based on the analysis of the soybean progress information and satellite images for the last 14 years. Moreover, this period was selected in order to get images covering the time series when the soybean reflectance and yield have the highest correlation (Johnson, 2014).

Two weather variables were selected to be evaluated on the models: daytime land surface temperature (LST), and precipitation. The LST is a similar, but not exactly the same, measurement as more commonly collected air temperature. The two variables (LST and air temperature) are strongly related, though, with LST having larger temperature extremes and being locally dependent on the land cover type (Mildrexler et al., 2011; Wan, 2008). The LST was produced from the 8-day composited thermal product from Aqua satellite's MODIS sensor (termed MYD11A2). Daily precipitation data was provided by the Climate Hazards Group Infrared Precipitation with Stations (CHIRPS) dataset. The CHIRPS provides precipitation data at ~5.5 km resolution by merging satellite and weather station information. This source of data (CHIRPS) uses satellite in three ways: first, satellite means are used to produce high-resolution rainfall climatologies; second infrared Cold Cloud Duration fields are used to estimate daily rainfall deviation from climatologies. Lastly, satellite precipitation fields are used to guide interpolation through local distance decay functions (Cunha et al., 2018). Precipitation layers were re-projected and down-scaled in order to be combined with the rest of the collected data. Precipitation was accumulated (summed) in an 8 days period to match with NDVI and LST derived from MODIS.

Data collection and organization

Since Brazil does not have a crop-specific data layer for retrieving geographical information about soybean field locations, we decided to use the data from CAR. The CAR is

an electronic national public registry, mandatory for all rural properties, with the purpose of integrating the environmental information related to the permanent preservation areas (restricted use), remnants of forests, other forms of native vegetation, and the consolidated areas, composing a database for control, monitoring, environmental and economic planning against deforestation. For the purposes of this study, we selected the consolidated rural areas, that is considered as an area of rural property with anthropogenic occupation preexisting on July 22, 2008. This information was downloaded as individual shapefiles (one for each municipality considered in this study), and then merged via R (R Core Team, 2017) in a unique file to be uploaded on the GEE platform.

All the VIs and the weather data were gathered via GEE using the CAR layer as a cropland mask. All the collected information was organized in a table format and averaged to municipality level before being merged with the yield data layer, comprising the first and the second steps on the model development (Figure 2).

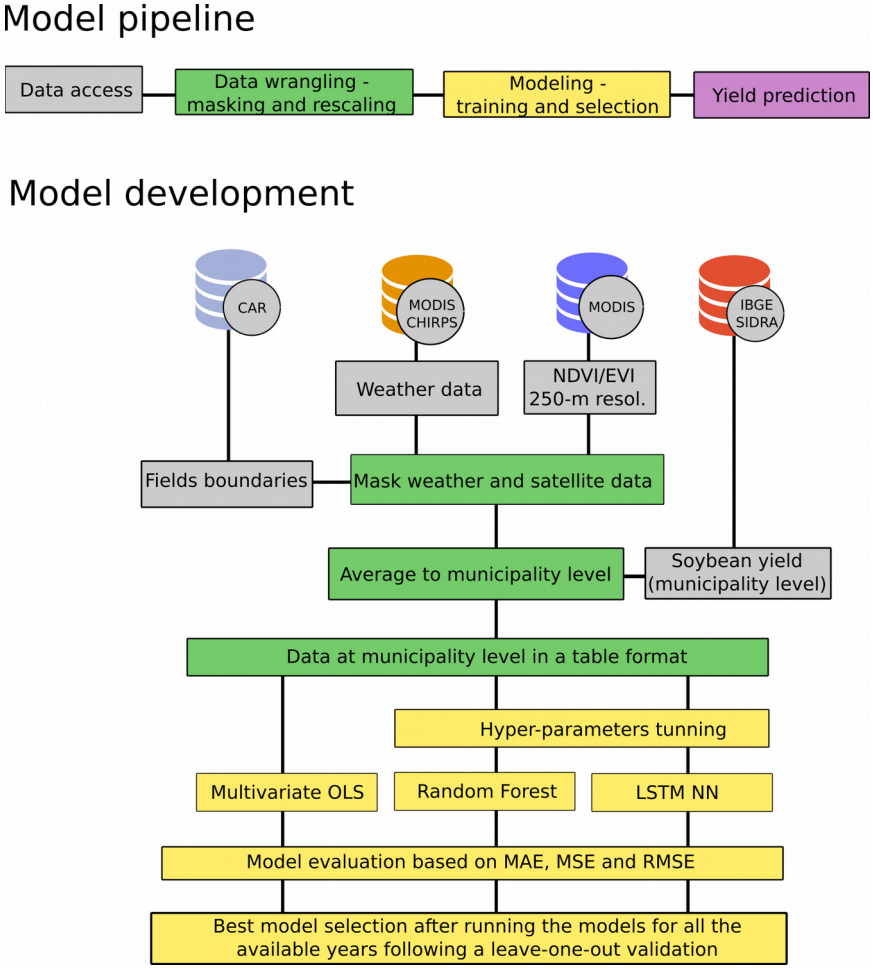


Figure 2. Flowchart indicating all steps of the model development: 1- data access, 2- data wrangling which includes masking gridded data using CAR field boundaries and re-scaling

the satellite and weather data to municipality-level before merging it with the yield data, and step 3- building the empirical relationships between soybean yield and the predictors (enhanced vegetation index - EVI, normalized difference vegetation index - NDVI, land surface temperature - LST, and precipitation) for the three considered algorithms (multivariate OLS, random forest, LSTM neural network), and selecting the best model based on metrics (MAE, MSE, and RMSE) derived from a leave-one-year-out cross validation.

Empirical relationships between yield, vegetation indices and weather

Three algorithms were tested to describe the relationship between yield, VIs and weather: i) multivariate OLS linear regression, ii) random forest, and iii) LSTM neural network. Multivariate OLS model was choosen as a benchmark relative to the two machine learning algorithms, since it represents the one of the simplest form to build empirical relationships between dependent and independent variables. Secondly, we chose the random forest model to explore non-linear models. Random forests are easy to train relaying on tuning only two hyper-parameters, the number of variables in the random subset at each node and the number of trees in the forest, and the output is usually not very sensitive to their values, avoiding any subjectivity (Liaw and Wiener, 2002). In addition, they have low sensibility to outliers, resulting in high computational efficiency and robustness against over-fitting (Belgiu and Drãgut, 2016). Lastly, we tested the model performance using the LSTM neural network. The LSTM neural network are prepared for receiving sequential data as an input and are able to extract important aspects related to the time series since it maintains a chain structure with time steps, similar to the way that crop growth modeling works. Each step takes information from previous step and outside input (from feature space – new NDVI, EVI, LST and precipitation values), and provides output for the next step. Furthermore, during the training process this algorithm is capable of retaining key information of input signals, and ignore less important parts.

For multivariate OLS and random forest, two classes of predictors were tested: i) the multi temporal EVI, NDVI, LST and precipitation, and ii) the seasonal integrated EVI, NDVI, LST and precipitation (as cumulative over the growing season). Therefore, for those two algorithms the annual municipality-level soybean yield forecasting model can be written as the following function:

$$y_{ij} = f(x_{ij}) + e_{ij} \quad (1)$$

where, y_{ij} is soybean yield for the i_{th} municipality and j_{th} year, x is the user-selected vector of predictors, f is a user-selected computer algorithm, and e_{ij} is error associated with the prediction.

The LSTM neural network received the two classes of inputs at the same time, classified as dynamic and static data. The dynamic data were related to the VIs and weather time series, and were organized in a 3D array (samples, time steps, and features). The static data were the seasonal integrated variables. A concatenated layer was used to deal with those different input dimensions.

Since random forest and LSTM neural network are machine learning algorithms, there is a need for defining some hyper-parameters (parameters that the algorithm cannot lean from the data). For random forest the considered hyper-parameter were the number of variables in the random subset at each node and the number of trees in the forest. For the LSTM neural network, we tuned the number of hidden layers, number of neuron on each hidden layer, dropout rate, batch size, activation function, learning rate, learning rate decay, and the gradient descent optimization algorithms. Moreover, the number of epoch was set to 60 and the training made use of the *EarlyStopping* callback function from the Keras (Chollet, 2015), with a patience parameter (the number of epochs with no improvement after which training is stopped) equal to 20 to avoid over-fitting. Four years were randomly selected from the data: 2009, 2010, 2012 and 2016 for fine-tuning the machine learning hyper-parameters (sensitivity analyses showed that the changes in the selected years did not significantly impact on the model parameterization). We performed a random search in order to find the best values for the hyper-parameters for the two considered algorithms.

For all the algorithms, model performance was evaluated using a leave-one-year-out cross-validation approach and three metrics were used to assess the model accuracy: the mean absolute error (MAE), the mean square error (MSE) and the root-mean square error (RMSE). The MAE represents the average magnitude of the errors while RMSE is a quadratic scoring rule for the average magnitude of the error, and it is more useful when large errors are particularly undesirable. The RMSE will always be larger or equal to the MAE; the greater difference between them, the greater the variance in the individual errors in the sample. If the RMSE equals to the MAE, then all the errors are of the same magnitude. The MSE was chosen because this metric can be dissected into two components, $bias^2$ (squared bias) and variance (σ^2), and this decomposition is helpful to understand if the model error has a more systematic or non-systematic structure.

Time series sensitivity analysis

For all the models, a sensitivity analysis was performed to check how early in the crop growing season the forecasting yield model can be implemented and its impact on the overall model performance. For this purpose, data collected later during the growing season was subsequently removed from the model and the same validation approach aforementioned was used to compute the MAE, MSE, and RMSE. Thus, we tested the models using data until DOY 16 (January 16), DOY 32 (February 1), DOY 48 (February 17), and DOY 64 (March 5). We have assumed the existence of a delay in the release of the yield forecast models based on the process for uploading the MODIS product by NASA, in approximately five days (Sakamoto et al., 2014).

The model training highlighted in the step 3 of the model development framework (Figure 2) was performed in the R environment using the RandomForest (Liaw and Wiener, 2002) and the Keras (Chollet, 2015) packages.

Relationship between model accuracy and yield/weather anomalies

Long-term yield data (1972-2017) for the entire region considered in this study (average over all the municipalities) was collected from IBGE. A regression analysis was performed using year as the independent variable and yield as the response variable. The residuals from this relationship (yield anomalies) were used in a Monte Carlo simulation in R program aiming at estimating the likelihood of any particular event to occur. We assume that the yield anomalies follow a normal distribution with mean and standard deviation estimated from the data. Residuals from the fitted model were utilized instead of using the absolute yield value to account for the genetic and technological evolution over the years.

We repeated this task using weather data instead of yield, and for doing that we extracted long-term (1982-2018) temperature and precipitation information from NASA POWER for all the municipalities considered in this study. We used NASA POWER for this analysis instead of MODIS and CHIRPS because MODIS only has information available after 2000. This information was summarized in 8-days periods (average for temperature and sum for precipitation). A Pearson correlation was performed among all the 8-days periods for precipitation and temperature, and yield in order to find a contiguous period of high correlation between these weather variables and yield. After defining this period, precipitation and temperature were summarized for the entire period and a Monte Carlo simulation was performed assuming that precipitation and temperature follow a multivariate normal distribution with μ_1, μ_2 and Σ, where: μ_1 is the precipitation mean, μ_2 is the temperature mean

and Σ the variance-covariance matrix between precipitation and temperature. We decide to use a bivariate normal distribution instead a high dimensional distribution to avoid problems related to the curse of dimensionality, when the dimension is large and the sample size is moderate (Amato et al., 2013).

Results

Model performance at different forecast dates

Regardless of the date of the forecast, the seasonal integrated predictors outperformed the multi-temporal ones for multivariate OLS regression and random forest (data not shown). As the soybean yield forecasts were performed earlier in the growing season all the models tended to become less accurate. Overall, the LSTM neural network presented the lowest values for MAE, MSE, and RMSE compared to the rest of the tested models, except for DOY 16 where the LSTM had the least accurate performance among the three options, with the best performance for the multivariate OLS (Table 1).

The observed versus predicted soybean yield for the four dates tested in our model were explored using the best algorithm for each specific date. Based on the data presented on Table 1, we used the multivariate OLS regression model for DOY 16 and the LSTM for the remaining dates (Figure 3A-D). The overall soybean yield data distribution for RS, Brazil from 2003 to 2016 presented a wide range of values from 0.2 to 4.2 Mg ha^{-1} with no evidence to reject the null hypothesis that the sampled yield values came from a normally distributed population (Shapiro-Wilk test p-value>0.05). The maximum likelihood estimation for the mean and standard deviation based on the data were 2.4 and 0.8 Mg ha^{-1} respectively.

Table 1. Model metrics comparison among multivariate OLS, random forest, and LSTM neural network.

Day of	MAE (Mg ha^{-1})			RMSE (Mg ha^{-1})			MSE (kg ha^{-1})^2		
year	OLS	RF	LSTM	OLS	RF	LSTM	OLS	RF	LSTM
DOY16	0.42	0.46	0.52	0.53	0.57	0.68	0.28	0.33	0.46
DOY32	0.46	0.44	0.42	0.58	0.57	0.56	0.34	0.33	0.31
DOY48	0.40	0.37	0.25	0.50	0.48	0.32	0.25	0.23	0.10
DOY64	0.32	0.32	0.24	0.40	0.39	0.32	0.16	0.15	0.10

* Values presented for OLS (multivariate OLS regression) and RF (random forest) are related to models using the seasonal integrated variables.

Despite residuals have been equally distributed along the 1:1 line considering all the years together for the predicted versus observed yield models, this pattern was not followed when the years were analyzed individually. Years such as 2004 and 2005 presented an error greater than the others, mainly for the early season forecasts (DOY 16 and 32) (Figure 3). Moreover, after decomposing the MSE into its two components, the bias^2 and σ^2, it can be seen that for the years presenting a greater MSE, the highest contributions came from the bias^2 (lack of the capacity of the model to describe a specific phenomenon, systematic error) and not from σ^2 (non-systematic source of error) (Figure 3).

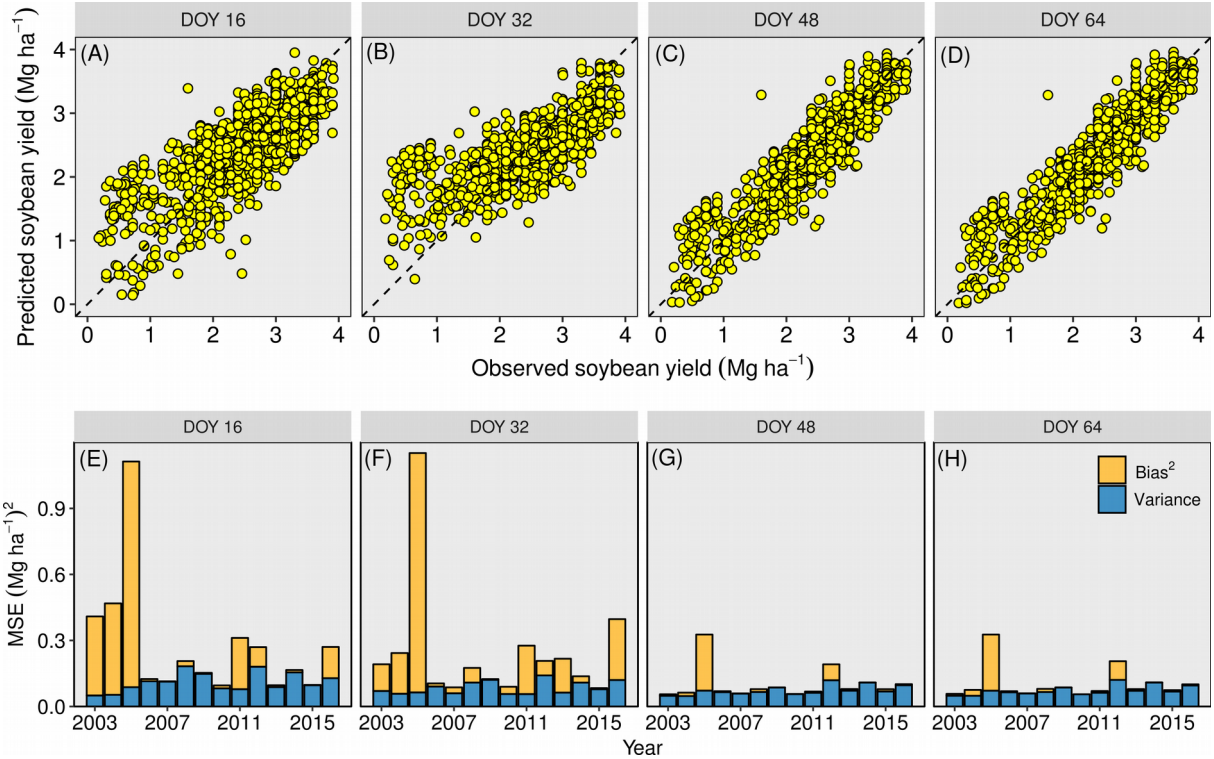


Figure 3. Upper panels (A to D) portraying the observed versus out-of-sample forecasted corn yield (forecast model with multi-temporal vegetation indices (VIs), land surface temperature and precipitation) for different dates expressed in days of year (DOY). A black dashed line portrays the 1:1 line for the predicted-observed relationship. The Long-Short Term Memory (LSTM) Neural Network used for DOY 32, 48 and 64. Multivariate OLS regression for DOY 16. In bottom panels (E to H) variations in the mean square error (MSE) and its decomposition in bias^2 and variance along the years for different dates expressed in DOY.

We calculated the cumulative probability frequency for the soybean yield anomalies (residuals from the soybean yield-year relationship) for the region considered in this study (Figure 4 A-C). The analyses showed that years presenting the greatest anomalies tended to present the highest MSE values, and consequently the highest values for $bias^2$ (Figure 4D). Moreover, it was demonstrated that the frequency of occurrence of years with anomalies equal or higher than the one found in 2005 year seems to be really negligible, ~0.7% or in other words 1 in ~142 years. Following a similar approach, but using weather data instead of yield, we built a second probability density function based on temperature and precipitation. For the second approach, we focused on a specific period of the soybean growing season in Brazil - between DOY 360 and DOY 56 (usually from flowering to seed filling stages), where these variables presented the highest correlation with yield (Figure 4E). Using this second approach the probability of occurrence for a year with an anomaly equal or higher to 2005 year was 0.3%, close to the 0.7% (but even smaller) that we estimated using the first approach.

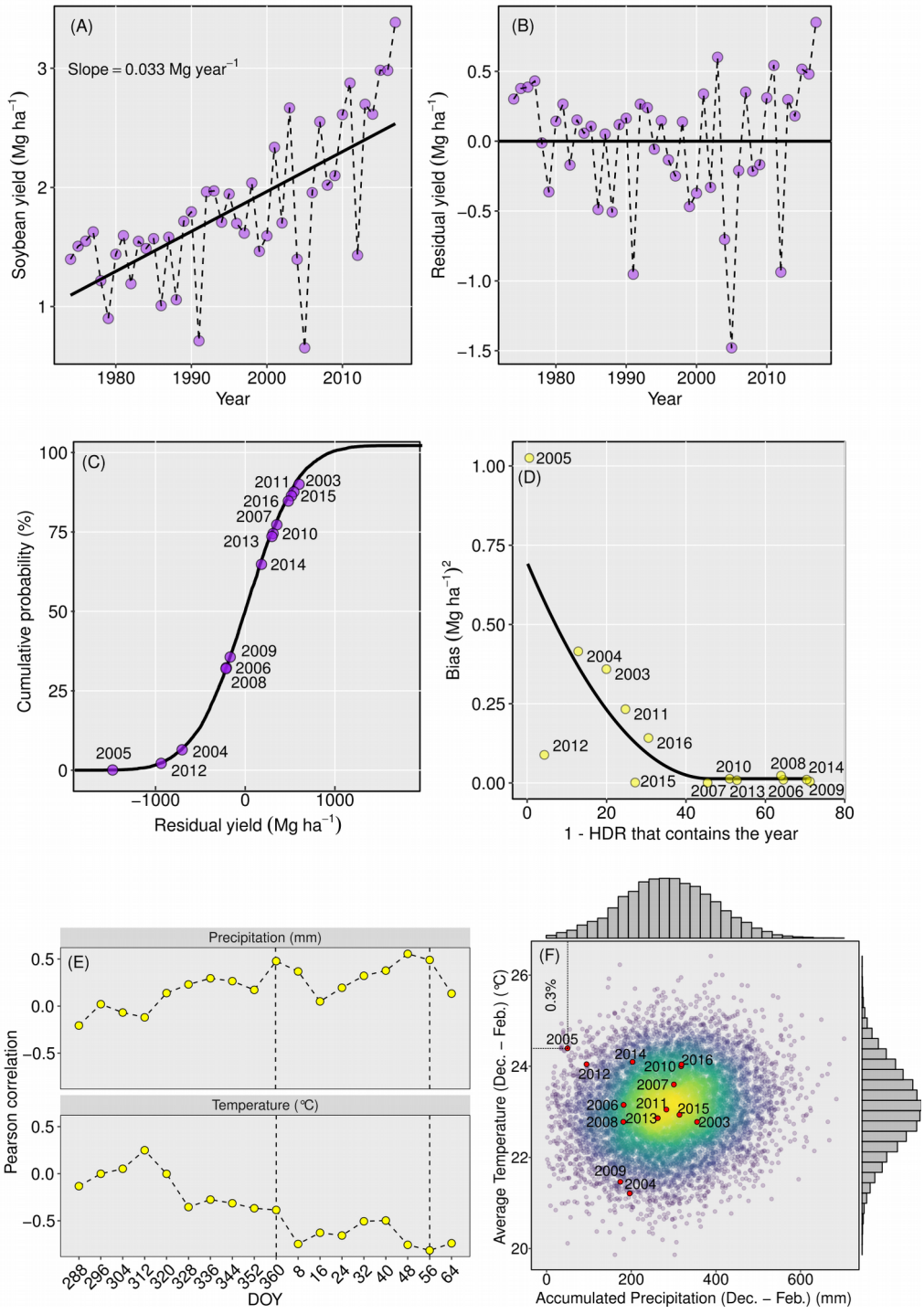


Figure 4. (A) Relationship between soybean yield and years for the study region. (B) Relationship between residuals for panel A and growing season year. (C) Cumulative distribution function estimated through the Monte Carlo simulation for the yield residuals. (D) Relationship between the Bias^2 from the DOY 16 yield forecast and the 1-High Density Region (HDR) needed to overlap the considered year – 1-HDR measures how far a specific year is from the mean of the distribution towards the tails, putting equal weights for both tails.

(E) Pearson's correlation between soybean yield and average air temperature and precipitation for different 8-days periods during the soybean growing season. (F) Multi-Gaussian probability density function estimated through the Monte Carlo simulation for average air temperature and precipitation (from DOY 360 to DOY 56) for the study region.

Discussion

Our results clearly showed that satellite imaging combined with weather data can provide useful information to develop more accurate models to forecast yields of soybean in Brazil. Crop yield forecast based on satellite imagery have become a popular tool for providing near real-time prediction of crop status from small (field and sub-field conditions) (Azzari et al., 2016; Jin et al., 2019, 2017; Lobell et al., 2015; Peralta et al., 2016; Schwalbert et al., 2018) to medium/large domains (county/state) (Bolton and Friedl, 2013; Cai et al., 2018; Johnson, 2014; Lobell, 2013; Peng et al., 2018; Sakamoto et al., 2014; Shao et al., 2015). Furthermore, the integration of canopy reflectance (sometimes summarized as VIs) and weather variables, have been demonstrated as a promising approach to enhance performance of yield forecast models. The negative correlation of heat, vapor pressure deficit, and the positive correlation of precipitation (Cai et al., 2018; Johnson, 2014; Peng et al., 2018) have been successfully explored in combination with multi-temporal VIs for providing more accurate near real-time forecasts for different crops.

Most of the algorithms used for exploring relationships between yield - multi-temporal VIs and weather variables rely on multivariate OLS (Cai et al., 2018; Lobell et al., 2015; Sakamoto et al., 2014), random forest (Cai et al., 2018; Shao et al., 2015), Rulequest Cubist, (Johnson, 2014), or supported vector machine (Cai et al., 2018). Despite those algorithms usually presents a satisfactory performance for the aforementioned task, they are not prepared for dealing with time-ordered data. Since VIs and weather variables are inherently temporal, with past state of these variables usually presenting on the future cause-effect relationship, algorithms able of learning patterns based on the sequence how the data is collected have a great potential for outperforming algorithms that treat data in a static viewpoint. In our study, the LSTM neural network outperformed the multivariate OLS regression and random forest for all the tested dates except for the earliest one. For the earliest date, there was less information from the past (related to the forecast date) to be learned by the LSTM neural network model. The use of LSTM for forecasting crop yield is still limited on literature with

only a few research studies exploring this topic (Cunha et al., 2018; Wang et al., 2018; You et al., 2017).

Regardless the choice of the algorithm for modeling the yield-predictors empirical relationship, one of the main challenges on using satellite and weather data as proxies to yield at a regional level still remain on the crop field detection, mainly for countries where the crop field boundary and crop-specific layers are not available. The main outcome of this research was a soybean yield forecast model able to predict yield at the municipality level in RS state, southern Brazil. This model has proven to present a high accuracy even without using any crop specific layer, with performance comparable to the models developed in the US by Johnson (2014) using the CDL as crop mask layer and You et al. (2017) using a general world-wide land cover data derived from MODIS (DAAC, 2015), and models developed in Brazil (for four municipalities in Paraná state), by Figueiredo et al. (2016). Similar results also have been reported for corn in the US, demonstrating that models based on multi-temporal NDVI summary statistics had similar performance either using a specific or general (e.g. summer crops, cultivated crop) crop masks (Shao et al., 2015). It is important to note that in the US Midwest and in RS state a corn-soybean rotation on an annual basis is widely adopted. More importantly, previous studies have shown that corn and soybean have relatively similar NDVI profiles (Shao et al., 2010; Wardlow and Egbert, 2008). Therefore, the inclusion of corn in the summer crop mask may still mimic the reflectance signal derived for soybean field only. In RS, the soybean/corn cultivated area is more towards to the soybean side (more frequency of this crop in the rotation), therefore most of the pixels included in this analysis came from soybean fields. The results presented in this paper represents a great prospect for providing municipality-level soybean yield data in a near real-time basis, contrasting with the frequency of the data currently released by SIDRA/IBGE, with the last yield estimation (2016/2017 growing season) announced in 2018.

Furthermore, we extended our analysis pursuing to explore the sensitivity of the time for the forecast model, considering that the importance of a yield forecast is a balance between its accuracy and the timing when the prediction is performed, and usually there is a trade-off between the error and the date of the prediction (Bolton and Friedl, 2013; Sakamoto et al., 2014; Shao et al., 2015; You et al., 2017). Our results clearly reflected this trade-off since as the forecast is anticipated during the growing season the error of the model tended to rise. Despite of that, soybean yield still can be forecasted at municipality-level in RS, Brazil at DOY 16 with a MAE of 0.42 Mg ha^{1}, and a RMSE of 0.53 Mg ha^{-1}. The penalization in model accuracy for anticipating the yield forecast was greater for years with extreme weather

(anomalies from the normal weather) but most of the error from the MSE came from $bias^2$ instead of σ^2. The latter shows that even for years with conditions highly adverse, the model was still able to predict the most and least yielding municipalities even without accurately predicting the absolute soybean yields.

Moreover, yield anomalies such as the ones reported in the 2005 soybean growing season in southern Brazil are unlikely to happen, and the reported model performance (RMSE, MSE, and MAE) was highly penalized by the errors associated with this growing season. After dissecting MSE in σ^2 and $bias^2$ for each one of the years, it became quite clear that years with a lower probability to occur had the highest $bias^2$, and the $bias^2$ tended to decrease and get stable as the years were settled towards the middle of the yield anomalies distribution (high-density region). The relationship between the probability of a specific type of year to occur and the $bias^2$ is in fact related to the lack of information about that event in the training dataset. Future applications of this model under conditions similar to 2005 year are expected to result in accurate soybean yield forecast, because those events (weather variation) will be already present on the training data.

Conclusions

Multi-temporal satellite imagery combined with weather variables can provide useful information, allowing the development of more precise yield forecast models to monitor soybean yield at municipality level. A decrease in the accuracy of the yield forecast model is expected by anticipating the date for yield prediction before harvest, but this study suggests that soybean yield can be predicted by DOY 16 (January 16) with reasonable accuracy. This is approximately 70 days before harvest in RS. Better accuracy (MAE of 0.24 Mg ha^{-1}) can be obtained by DOY 48 (February 17) - 40 days before harvest in RS. The LSTM neural network has been tested to have a better performance relative to random forest or the multivariate OLS regressions, mainly for predictions towards the end of the growing season plausible due to the amount of data collected to compose the time series.

The training and validation approaches were adequate to test the model performance in different weather and yield conditions. Model performance for years with more adverse weather conditions (dramatically different from the normal years) and consequently with higher yield anomalies related to the historical yield distribution is expected to be inferior compared to the overall model accuracy for the remaining years. Under extreme weather conditions, the increase in the error was mainly associated with $bias^2$ than σ^2. For this reason,

we expect an increase in the model generalization for future extreme weather events as more data is added into the training process. Despite the analysis being developed for southern Brazil, the general approach described in this study can be potentially applied to other geographical regions around the globe with similar availability of data. This could contribute to support agricultural decisions in regard to managing and transferring risks within crop production and to improve overall crop predictions for policy makers.

Acknowledgments: This study was supported by CAPES Foundation, Ministry of Education of Brazil, Brasilia - DF, Zip Code 70.040-020, process 88887.130848/2016-00. Contribution no. KAES no. 19-XXX-J from the Kansas Agricultural Experiment Station.

References

Alvarez, R., 2009. Predicting average regional yield and production of wheat in the Argentine Pampas by an artificial neural network approach. Eur. J. Agron. 30, 70–77. https://doi.org/10.1016/j.eja.2008.07.005

Amato, U., Antoniadis, A., Carfora, M.F., Colandrea, P., Cuomo, V., Franzese, M., Pignatti, S., Serio, C., 2013. Statistical classification for assessing prisma hyperspectral potential for agricultural land use. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 6, 615–625. https://doi.org/10.1109/JSTARS.2013.2255981

Azzari, G., Jain, M., Lobell, D.B., 2016. Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries. Remote Sens. Environ. 202, 129–141. https://doi.org/10.1016/j.rse.2017.04.014

Belgiu, M., Drãgut, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS J. Photogramm. Remote Sens. 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Bolton, D.K., Friedl, M.A., 2013. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. Agric. For. Meteorol. 173, 74–84. https://doi.org/10.1016/j.agrformet.2013.01.007

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1007/9781441993267_5

Cai, Y., Guan, K., Lobell, D.B., Potgieter, A.B., Wang, S.W., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., 2018. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. Am. Geophys. Union, Fall Gen. Assem. 274, 144–159. https://doi.org/10.1016/j.agrformet.2019.03.010

Chollet, F., et al., “Keras,” https://github.com/fchollet/keras, 2015.

Cunha, R.L.F., Silva, B., Netto, M.A.S., 2018. A scalable machine learning system for pre-season agriculture yield forecast. Proc. - IEEE 14th Int. Conf. eScience, e-Science 2018 423–430. https://doi.org/10.1109/eScience.2018.00131

DAAC, N.L., 2015. The MODIS land products. URL http://lpdaac.usgs.gov

Embrapa, 2018. Soja em números. URL https://www.embrapa.br/soja/cultivos/soja1/dados-economicos

Ferencz, C., Bognár, P., Lichtenberger, J., Hamar, D., Tarcsai, G., Timár, G., Molnár, G., Pásztor, S., Steinbach, P., Székely, B., Ferencz, O.E., Ferencz-Árkos, I., 2004. Crop yield estimation by satellite remote sensing. Int. J. Remote Sens. 25, 4113–4149. https://doi.org/10.1080/01431160410001698870

Figueiredo, G.K.D.A., Brunsell, N.A., Higa, B.H., Rocha, J.V., Lamparelli, R.A.C., 2016. Correlation maps to assess soybean yield from EVI data in Paraná State, Brazil. Sci. Agric. 73, 462–470. https://doi.org/10.1590/0103-9016-2015-0215

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Hamada, Y., Ssegane, H., Negri, M.C., 2015. Mapping intra-field yield variation using high resolution satellite imagery to integrate bioenergy and environmental stewardship in an agricultural watershed. Remote Sens. 7, 9753–9768. https://doi.org/10.3390/rs70809753

Jin, Z., Azzari, G., Lobell, D.B., 2017. Improving the accuracy of satellite-based high-resolution yield estimation: A test of multiple scalable approaches. Agric. For. Meteorol. 247, 207–220. https://doi.org/10.1016/j.agrformet.2017.08.001

Jin, Z., Azzari, G., You, C., Di Tommaso, S., Aston, S., Burke, M., Lobell, D.B., 2019. Smallholder maize area and yield mapping at national scales with Google Earth Engine. Remote Sens. Environ. 228, 115–128. https://doi.org/10.1016/j.rse.2019.04.016

Johnson, D.M., 2014. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. Remote Sens. Environ. 141, 116–128. https://doi.org/10.1016/j.rse.2013.10.027

Johnson, D.M., Mueller, R., 2010. The 2009 Cropland Data Layer. Photogramm. Eng. Remote Sens. 76, 1201–1205.

Johnson, M.D., Hsieh, W.W., Cannon, A.J., Davidson, A., Bédard, F., 2016. Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. Agric. For. Meteorol. 218–219, 74–84. https://doi.org/10.1016/j.agrformet.2015.11.003

Khaki, S., Wang, L., 2019. Crop Yield Prediction Using Deep Neural Networks.

Li, A., Liang, S., Wang, A., Qin, J., 2013. Estimating Crop Yield from Multi-temporal Satellite Data Using Multivariate Regression and Neural Network Techniques. Photogramm. Eng. Remote Sens. 73, 1149–1157. https://doi.org/10.14358/pers.73.10.1149

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2, 18–22.

Lobell, D.B., 2013. The use of satellite data for crop yield gap analysis. Field Crop. Res. 143, 56–64. https://doi.org/10.1016/j.fcr.2012.08.008

Lobell, D.B., Thau, D., Seifert, C., Engle, E., Little, B., 2015. A scalable satellite-based crop yield mapper. Remote Sens. Environ. 164, 324–333. https://doi.org/10.1016/j.rse.2015.04.021

Mildrexler, D.J., Zhao, M., Running, S.W., 2011. A global comparison between station air temperatures and MODIS land surface temperatures reveals the cooling role of forests. J. Geophys. Res. Biogeosciences 116, 1–15. https://doi.org/10.1029/2010JG001486

Peng, B., Guan, K., Pan, M., Li, Y., 2018. Benefits of Seasonal Climate Prediction and Satellite Data for Forecasting U.S. Maize Yield. Geophys. Res. Lett. 45, 9662–9671. https://doi.org/10.1029/2018GL079291

Peralta, N., Assefa, Y., Du, J., Barden, C., Ciampitti, I., 2016. Mid-season high-resolution satellite imagery for forecasting site-specific corn yield. Remote Sens. 8, 1–16. https://doi.org/10.3390/rs8100848

R Core Team, 2017. R: A Language and Environment for Statistical Computing.

S. T. Drummond, K. A. Sudduth, A. Joshi, S. J. Birrell, N. R. Kitchen, 2013. Statistical and Neural Methods for Site-Specific Yield Prediction. Trans. ASAE 46, 1–10. https://doi.org/10.13031/2013.12541

Sakamoto, T., Gitelson, A.A., Arkebauer, T.J., 2014. Near real-time prediction of U.S. corn yields based on time-series MODIS data. Remote Sens. Environ. 147, 219–231. https://doi.org/10.1016/j.rse.2014.03.008

Schwalbert, R.A., Amado, T.J.C., Nieto, L., Varela, S., Corassa, G.M., Horbe, T.A.N., Rice, C.W., Peralta, N.R., Ciampitti, I.A., 2018. Forecasting maize yield at field scale based on high-resolution satellite imagery. Biosyst. Eng. 171, 179–192. https://doi.org/10.1016/j.biosystemseng.2018.04.020

Shao, Y., Lunetta, R.S., Ediriwickrema, J. and Iiames, J., 2010. Mapping Cropland and Major Crop Types across the Great Lakes Basin using MODIS-NDVI Data. Photogramm. Eng. Remote Sensing 75, 73–84. https://doi.org/10.14358/PERS.76.1.73

Shao, Y., Campbell, J.B., Taff, G.N., Zheng, B., 2015. An analysis of cropland mask choice and ancillary data for annual corn yield forecasting using MODIS data. Int. J. Appl. Earth Obs. Geoinf. 38, 78–87. https://doi.org/10.1016/j.jag.2014.12.017

USDA, 2019. USDA Foreign Agricultural Service. URL https://www.fas.usda.gov/regions/brazil

Wan, Z., 2008. New refinements and validation of the collection-6 MODIS land-surface temperature/emissivity product. Remote Sens. Environ. 140, 36–45. https://doi.org/10.1016/j.rse.2013.08.027

Wang, A.X., Tran, C., Desai, N., Lobell, D., Ermon, S., 2018. Deep Transfer Learning for Crop Yield Prediction with Remote Sensing Data 1–5. https://doi.org/10.1145/3209811.3212707

Wardlow, B.D., Egbert, S.L., 2008. Large-area crop mapping using time-series MODIS 250 m NDVI data: An assessment for the U.S. Central Great Plains. Remote Sens. Environ. 112, 1096–1116. https://doi.org/10.1016/j.rse.2007.07.019

You, J., Li, X., Low, M., Lobell, D., Ermon, S., 2017. Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data. Proc. Thirty-First AAAI Conf. Artif. Intell. 4559–4566. https://doi.org/10.1109/MWSCAS.2006.381794

4 ARTIGO 3 – FORECASTING MAIZE YIELD AT FIELD SCALE BASED ON HIGH-RESOLUTION SATELLITE IMAGERY

Abstract: Estimating maize (*Zea mays* L.) yields at the field level is of great interest to farmers, service dealers, researchers, and policy-makers. The main objectives of this study were to: i) provide guidelines on data selection aimed at building forecasting yield models using Sentinel-2 satellite imagery; ii) compare different approaches and vegetation indices (VIs) during model building; and iii) perform spatial and temporal validation to see if empirical models could be applied to other regions or when models coefficients should be updated. Data analysis was divided into four major steps: i) data acquisition and preparation; ii) selection of training data; iii) building of forecasting yield models; and iv) spatial and temporal validation. The analysis were performed using yield data collected from 19 maize fields located in Brazil – Rio Grande do Sul state (2016/2017 season) and Mato Grosso state (2016 and 2017 seasons) – and in the United States – state of Kansas (2016 season), and VIs (NDVI, green NDVI and red edge NDVI) derived from Sentinel-2. Main outcomes from this study were: i) data selection impacted yield forecast model and fields with narrow yield variability and/or with skewed data distribution should be avoided; ii) models considering spatial correlation of residuals outperformed OLS regression; iii) red edge NDVI was most frequently retained into the model compared with the other indices; and iv) model prediction power was more sensitive to yield data frequency distribution than to the geographical distance or years. Thus, this study provided guidelines to build more accurate maize yield forecasting models, but also established limitations for up-scaling, from farm-level to county, district, and state-scales.

Keywords: forecasting yield models; maize; satellite imagery; yield maps; model validation; Sentinel-2

1. Introduction

Precise and reliable yield forecast tools could play a fundamental role in supporting policy formulation, and decision-making process in agriculture (e.g. storage and transport) (Córdoba et al., 2016; Kantanantha et al., 2010; Stone and Meinke, 2005). Historically, most models developed for yield forecasting are focused to large domains (between-field variability), (DiRienzo et al., 2000; Doraiswamy et al., 2003; Hamar et al., 1996; Lopresti et al., 2015; Reeves et al., 2005; Sibley et al., 2014), mostly because, in the past there was limited source of data with a sufficient temporal and spatial resolution for accurate within-field crop yield estimates. Nowadays, satellite data have become more accessible (Azzari, et al., 2017) with more options of high resolution imagery, such as Skysat RapidEye, and Sentinel-2 satellites, and more studies have portrayed the benefits of using high-resolution satellite imagery for identifying within-field yield variation (Azzari et al., 2017; Jin et al., 2017 ;Peralta et al., 2016). Among the high resolution satellites Sentinel-2, that is a joint

initiative of the European Commission (EC) and the European Space Agency (ESA), represents a great opportunity towards fine resolution yield forecast models since, it is publically accessible satellite and was design to provide systematic global acquisitions of high-resolution (10 to 20m) multi-spectral imagery with a high revisit frequency (5 days at equator) (Drusch, et al., 2017).

The potential to forecast yield using satellite information is already known and a wide set of statistical approaches have been explored. Some approaches rely on the statement that total biomass production is closed related to the fraction of photosynthetically active radiation absorbed by vegetation (fAPAR) over the course of the growing season (Monteith 1977). fAPAR estimations are most often derived from VIs (Lobel, 2013), since the linear relationships between those two variables are well-known (Myneni et al., 1994). However, considering that most remote sensing data are not available on a daily basis, some interpolation is needed to estimate daily fPAR and this task becomes a challenge with a low number of images.

Empirical relationships between ground-based yield measures and remote sensing data have been considered as the simplest approach to forecast yield with low computational power demanding (Hatfield et al., 2008; Lobell, 2013), and have been successfully implemented in several studies with maize (Bognár et al., 2011; Bu et al., 2017; Lobell et al., 2015; Peralta et al., 2016; J. Shanahan et al., 2001; Sibley et al., 2014). The separation of data into the training and validation datasets is a common practice allowing self-test model replicability irrespective of the difference between the two datasets in space or time. Selection of ground-truth data to build models is one of the most important steps aiming at getting reliable yield predictions, and it is known that nature and volume of data have a direct impact on the model quality (Hatfield et al., 2008; Schwalbert et al., 2018). Despite that mostly studies randomly selected a subset of the data for comprising training or validation data (Gholap et al., 2012; Gonzalez-Sanchez et al., 2014; Sheridan, 2013; Peralta et al., 2016; Yared et al., 2016) without any guideline.

Thus, aiming at model constructing, the choice of fields in order to get a representative sample is very important. Moreover, the choice of the statistical model used to forecast yield have a large impact on the final result (Anselin et al., 2004; Peralta et al., 2016). Mostly empirical yield forecasting models based on VIs utilize classical ordinary least squares (OLS)-based on simple or multiple regression techniques (Noureldin et al., 2013; Rembold et al., 2013; J. Shanahan et al., 2001), without properly accounting for the spatial autocorrelation structure evolving these variables (Imran, et al., 2013; Peralta et al., 2016). The latter situation

can lead to problems with inflated variance and likely resulting in wrong conclusions (Anselin et al., 2004; Bongiovanni et al., 2007).

A second constraint related to models derived from simple empirical relationships is that they tend to be time- and space-limited, valid only under similar conditions as when the correlation was established (Hatfield et al., 2008; Lobell, 2013; Tucker et al., 1980). Currently, the potential to forecast yield using satellite information through empirical models is already known, but the challenge is to extend these tools beyond the structured environment of research studies (Hatfield et al., 2008). Lastly, the selection of adequate VIs also an important step in model development (Peralta et al., 2016). The normalized difference vegetation index (NDVI) (Rouse et al., 1974) is one the most widely used VIs to assess crop growth and yield (Peralta et al., 2016; Raun et al., 2002; Rembold et al., 2013; Solie et al., 2012), and it becomes as somewhat of a benchmark for researchers developing new VIs (Hatfield et al., 2008). However, there are some constraints related to saturation in medium to high leaf area index (LAI) values with this VI (Tucker, 1979; Haboudane et al., 2004; Nguy-Robertson et al., 2012). Thus, the incorporation of other indices that still have sensibility in high values of LAI, such as green NDVI (NDVIG) (Gitelson et al., 1996) and red-edge NDVI (NDVIre) (Gitelson and Merzlyak, 1994), have been reported as an important technique to improve empirical models (Hatfield et al., 2008; Peralta et al., 2016).

Following this rationale, guidelines for implementing yield forecasting models derived from empirical relatioships and for validating their spatio-temporal relevancy still remain unknown. Thus, the objectives of this study were to: i) identify parameters to guide data selection aiming at building forecasting yield models using Sentinel-2 satellite imagery; ii) compare different approaches (OLS vs. spatial correlation) and different VIs during the model building process; iii) perform spatial and temporal model validation using independent datasets to identify potential limitations in up-scaling forecasting yield models. The main hypothesis is that model predictability power increases as the yield frequency distribution of the training data becomes more alike to the validation data even when considering diverse spatio-temporal scales (geographical distance or time, years).

2. Materials and Methods

The analysis was performed on end-season yield monitor data and mid-season. Sentinel-2 images image were collected during a critical period for determining the grain yield in maize (approximately 20 days before and 20 after flowering) (Johnson 2014; Sakamoto et al., 2014;

Peralta et al., 2016). Sentinel satellite imagery of selected maize fields in farm conditions located in Brazil (BR) (Figure 1A and 1B) and US (Figure 1C). Six fields from Rio Grande do Sul (RS) state (2016/2017 season) and seven fields from Mato Grosso (MT) state (five from 2016 season and two from 2017 season) were selected for comprising the BR database. The field size ranged from 20 to 130 ha. It is important to mention that for MT, fields were selected from the second season (mainly cultivated after the soybean) since the first season is harvested around February. Usually the during the second season in MT the maize yield is lower compare to RS due the less favorable weather condition. In RS the average temperature during the growing season is 20.4 ºC with an accumulated precipitation of 1080 mm, and in MT the average temperature during the second season is 23.7 ºC with an accumulated precipitation of 700 mm.

The United States (US) database was composed of six fields (2016 season), all located in the state of Kansas (KS). Kansas database was only considered as validation data in the last step (spatial validation) where the models previously build were used to forecast maize yield in Kansas fields, in oder to test our main hypotheses. Information related to harvest date, satellite imagery collection data, and specific coordinates (latitude, longitude) for each field were recorded (Table 1). Most of the BR fields were utilized for training purposes, comprising the training database. Fertilizer application rates, crop management, and tillage practices varied between fields.

Table 1. Descriptive information of maize yield and satellite data: state, season, geographical position, harvest date and imagery acquisition date.

Field	State	Season	Data	Latitude*	Longitude*	Harvest date	Imagery date
F1	RS	2016-2017	V	-28.48	-52.78	02/16/2017	11/29/2016
F2	RS	2016-2017	V	-28.53	-53.54	02/21/2017	11/29/2016
F3	RS	2016-2017	V	-28.18	-52.69	02/14/2017	11/29/2016
F4	RS	2016-2017	T	-28.32	-52.71	02/27/2017	11/29/2016
F5	RS	2016-2017	V	-27.62	-53.36	02/18/2017	11/29/2016
F6	RS	2016-2017	T	-28.53	-53.56	02/17/2017	11/29/2016
F7	MT	2016	V	-15.47	-54.01	07/02/2016	04/29/2016
F8	MT	2016	T	-15.57	-54.15	07/06/2016	04/29/2016
F9	MT	2016	V	-15.57	-54.16	07/05/2016	04/29/2016
F10	MT	2016	V	-15.56	-54.17	07/05/2016	04/29/2016
F11	MT	2016	T	-15.58	-54.15	06/30/2016	04/29/2016
F12	MT	2017	V	-15.15	-53.94	06/30/2017	04/24/2017
F13	MT	2017	V	-15.15	-53.94	06/29/2017	04/24/2017
K1	KS	2016	V	39.53	-97.21	09/27/2016	06/20/2016
K2	KS	2016	V	39.54	-97.15	10/01/2016	06/20/2016

K3	KS	2016	V	39.55	-97.22	10/03/2016	06/20/2016
K4	KS	2016	V	39.57	-97.23	09/30/2016	06/20/2016
K5	KS	2016	V	39.53	-97.23	09/22/2016	06/20/2016
K6	KS	2016	V	39.56	-97.24	09/29/2016	06/20/2016

*Decimal coordinates - WGS 84. RS = Rio Grande do Sul. MT = Mato Grosso. KS = Kansas. T = Training Database. V = Validation Database.

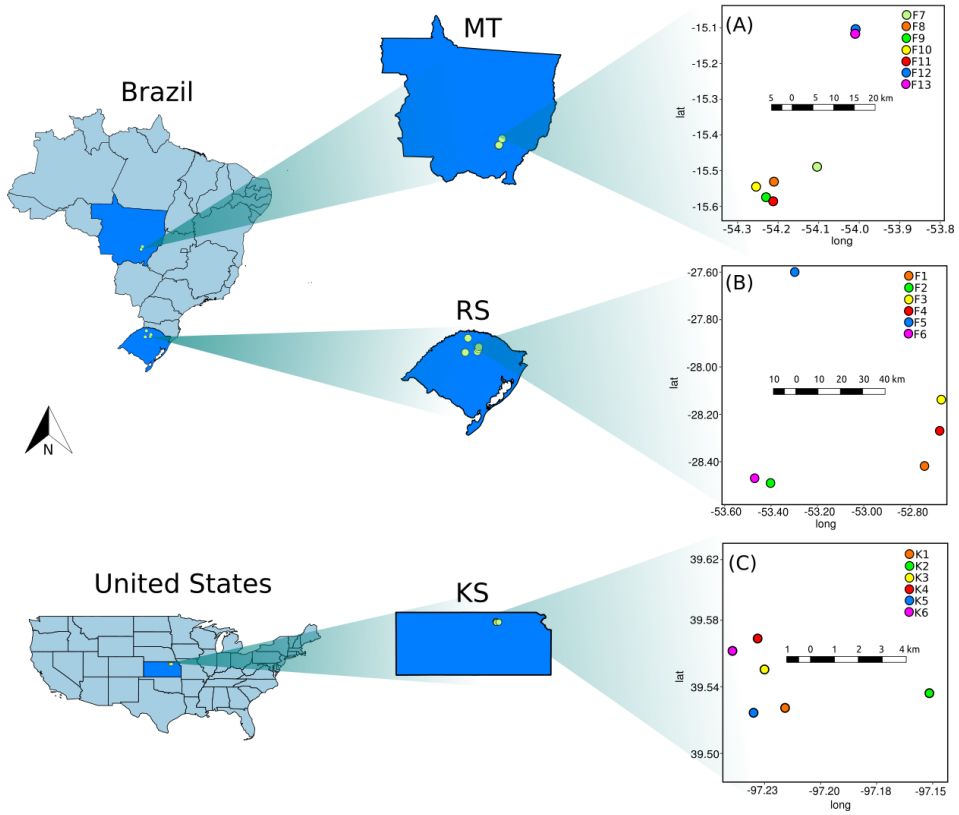


Figure 1. Field research studies located in Mato Grosso (MT) (A), Rio Grande do Sul (RS) (B), and Kansas (KS) (C). Circles represent the precise geo-position of the fields within each region. Scales bars are in different scales for panels A, B, and C.

This study was divided into four major steps, representing the main analysis performed to achieve the objectives (Figure 2). The four steps were: 1) data acquisition and preparation, 2) selection of training data, 3) building forecasting yield models, and 4) spatial and temporal validation (including fields from different growing season and geographies).

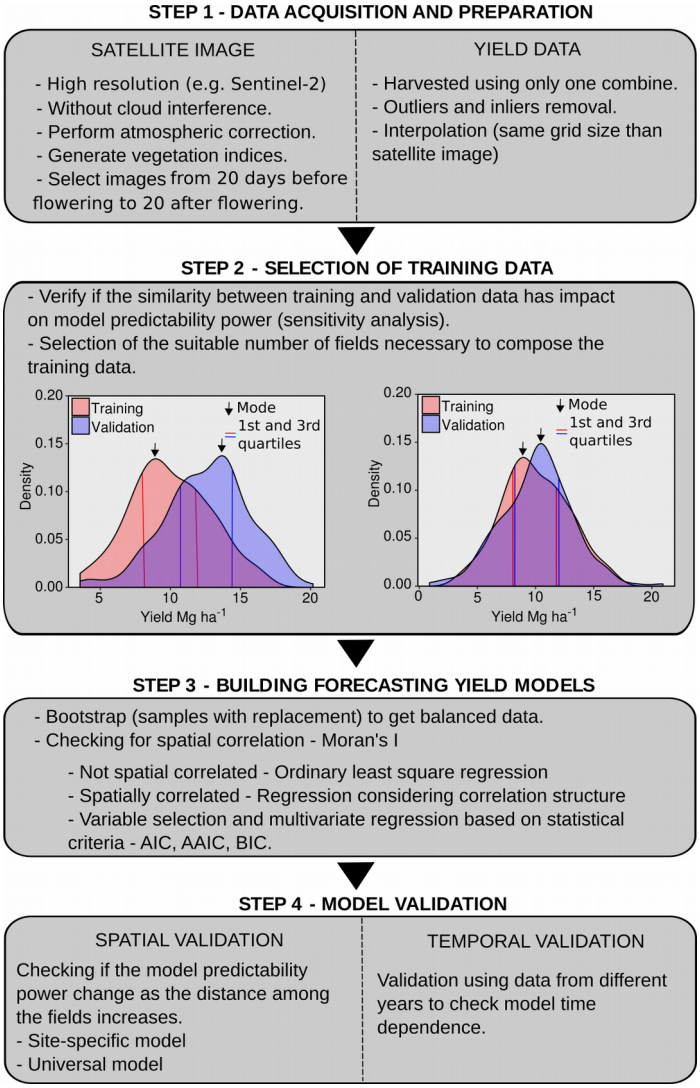


Figure 2. Theoretical framework indicating all steps of the analysis: step 1- data acquisition and preparation, step 2- selection of training data, step 3- building forecasting yield models, and step 4- model validation.

2.1. Data acquisition and preparation

The primary objective of this step was to establish criteria for selecting adequate quality of yield monitor (calibrated) and satellite imagery data. Yield data was submitted to a filter process in order to remove outliers and inliers. In this research, outliers were considered as values out of the mean ± 3 standard deviations (SD) range. According to Chebyshev's theorem (Amidan et al., 2005), it is inferred that a minimum of 89% of the data is within the mean ± 3 SD, regardless the data distribution. Inliers are data that differ significantly from their neighborhood but lie within the general range of variation of the data set (Córdoba et al., 2016). Spatial autocorrelation Moran's local index (Ii) (Anselin, 1995) was used to identifying

inliers. The Ii is basically applied individually to each neighborhood and shows the degree of similarity between an observation and its neighbors. In summary, *localmoran* function of the "spdep" R package (Bivand and Piras, 2015) was used to identify inliers. Moreover, the *moran.plot* function was implemented to calculate Ii and perform the Moran scatter plot to identify additional inliers. Further details can be found in Córdoba et al. (2016).

Spatial interpolation was performed to estimate maize yield values for areas where yield was not sampled. This procedure was required, even considering that yield monitor data was recorded in a high density (5 x 10 m), because after filtering yield density data was significantly decreased. Aiming at getting similar arrangement for all datasets, equivalent satellite imagery grid structure was used (10 x 10 m). Geostatistical interpolations involving semivariogram adjustment and ordinary kriging were performed, individually for each dataset, using R packages "geoR" (Ribeiro Jr and Diggle, 2016) and "gstat" (Pebesma, 2004).

Sentinel-2 images are composed by 10 bands with resolution between 10x10m and 20x20 m, in the visible, near infrared, and short wave infrared part of the spectrum. All bands were tested for its usefulness in building VI for yield forecast purposes. As a first step, a multivariate regression was applied to select the bands presenting greater correlation with yield; retaining only 6 bands, 3 (green), 4 (red), 5 (red-edge 1), 6 (red-edge 2), 8 (near-infrared), and 8a (red-edge 4) (Supplementary Table 1). The latter is in agreement with the scientific literature in the topic of forecasting crop yields using satellite data – primarily highlighting the importance of 5 bands (wavelengths): blue, green, red, red-edge, and near-infrared (Bu et al., 2017; DiRienzo et al., 2000; Doraiswamy et al., 2003; Hamar et al., 1996; Lobell et al., 2015; Lopresti et al., 2015; Peralta et al., 2016; Reeves et al., 2005; J. Shanahan et al., 2001; Sibley et al., 2014). The selected bands were employed to calculate 3 diverse VIs: NDVI, NDVIG, and NDVIre. The selection of the VI was based on previous researches showing the efficiency of these VI to forecast final maize yield (Bognár et al., 2011; Bu et al., 2017; Peralta et al., 2016; Shanahan et al., 2001). Sentinel-2 images were collected in a interval between 20 days before flowering and 20 days after flowering, depending on the availability of the image and the cloud interference (Table 1). Different satellite imagery data collection dates were tested for improved yield forecast, greater coefficient of determination (Figure S1). The red-edge band was resized to 10 m pixel size. Atmospheric correction was performed using the semi-automatic classification plugin in QGIS 2.18 (Congedo, 2016) in order to obtain surface reflectance without the interference of atmospheric gases. VIs, including NDVI, NDVIre, and NDVIG were generated using a combination of visible, near-infrared and red-edge bands.

2.2. Selection of training data

As previously detailed, since only selected BR fields (RS and MT) were used as training data; all KS fields were not used in this step. All fields were randomly sampled (bootstrap with replacement) to generate equal size of data points per field, 800 per field. Since one of the objectives of the paper was provided guidelines for training data selection three different alternatives of field selection were tested. Steps 2 and 3 in figure 2 were performed in a retroactive process for each one the data selection strategies.

The three data selection strategies tested to comprise the training data were: i) selection of the two fields with high yield amplitude and mostly recorded yield (more than 50% of the values) between first and third quartiles of the overall frequency distribution; ii) selection of the two fields with the lowest average yield among all fields (left shifted fields in relation to the overall distribution); and iii) selection of fields with the lowest, the highest and intermediate average yield among all fields (Supplementary table 2). For each one of the alternatives, the remained fields were considered as validation data. The aforementioned procedure was performed individually for RS and MT.

The similarity of training and validation distribution frequency was compared using two statistic parameters, mode and the Interquartile Range (IQR) position (range between the first and the third quartiles), skewness, and kurtosis. To compare the statistic parameters a 95% bootstrap percentile confidence interval (CI) (Efron and Tibshirani, 1993) was calculated using the “boot” package in R (Canty and Ripley, 2017), obtaining a total of 1000 bootstrap replicates to estimate the variability. Each time that training data was selected, models were build and validated with remain fields.

2.3. Building forecasting yield models

As aforementioned, this step occurred in parallel to the data selection step, forecasting yield models were built utilizing the selected training data to verify model predictability power and to access to the most important parameters driving to suitable training data selection.

As an initial phase, spatial autocorrelation analysis was conducted on yield and VIs (NDVI, NDVIG, and NDVIre) data of each field using Moran’s test. Moran’s I statistic measures the strength of spatial autocorrelation in a response among nearby locations in space as a function of cross-products of the neighboring weighted deviations from the mean.

Moran's I coefficient values near 1 and −1 indicate positive and negative autocorrelation, respectively. Coefficient near 0 refers to lack of spatial autocorrelation.

In order to identify an appropriate model that describes the relationship between end-season observed maize yields and VIs of mid-season imagery for the training data two approaches were considered. First, implementation of a linear regression model assuming that the errors are independent and identically distributed (i.i.d.). Ordinary least squares (OLS) method is known as an efficient procedure for estimating the unknown parameters for this model, herein termed as "OLS" model. When response (maize yield) and predictor variables (VIs), as well as the regression errors, exhibited spatial autocorrelation according to the Moran's I coefficient, the i.i.d. assumption was violated, and the application of models considering the spatial structure of the errors was pursued. Hence, models were adjusted using the *gls* function of the "nlme" R package (Pinheiro et al., 2017) with Gaussian, spherical and exponential spatial correlation of plotted errors.

At all steps above that require model selection, stepwise-regression procedure was used to determine the variables (VIs) that significantly contributed to yield prediction models. Stepwise forward was implemented using the function *stepAIC* of the "MASS" package (Venables and Ripley, 2002) from the R software. Statistical model comparison was performed using statistical criteria proposed by Akaike (AIC) (Johnson and Omland, 2004) and the coefficient of determination (R^2).

The multicollinearity (or collinearity) of the remaining bands was evaluated by computing the variance inflation factor (VIF). A threshold VIF value of 2 was established (Zuur et al., 2010) and a VI with VIF higher than 2 were removed from the model. The standardized coefficient was calculated using the R package "lm.beta" to check the weight of each VI into the model.

After running all the round for step 2 and step 3 (Figure 2) we checked for coincidences in similarities between training and validation data distribution according the parameters tested (mode, quartiles, skewness and kurtosis) and model accuracy assessing using RMSE (observed yield vs predicted yield).

Two categories of forecasting yield models were built: i) universal models, with both RS and MT training data and ii) site-specific models, for RS and MT states, obtaining one specific-model per state/region evaluated.

2.4. Spatial and temporal validation data

After selection of training data, a second validation was performed aiming at verifying spatial and temporal dependency on the models. For testing the first one, Kansas database was included as validation data. All the six sets of training data (three from RS and three from MT) were tested. The same approach discussed in the previous sections was applied. Yield frequency distribution of all the training data from RS and MT were compared with KS yield frequency distribution. After studying all yield frequency distributions, the most proper model was selected to forecast yield of the KS database (US), comprising six fields.

Temporal validation was performed using MT fields since only in MT there was data available from two different seasons (2016 and 2017). Forecasting yield model built using data from 2016 was used to estimate 2017 yields. The accuracy of estimation and model fitting was evaluated using the RMSE. In addition, spatial predictions from each model were visually compared with geostatistical interpolation of yield.

3. Results

3.1. Selection of training data

Different yield frequency distribution was documented for RS and MT. For RS, average maize yield was 12.7 Mg ha^{-1}, with 50% of the data (IQR) ranging from 10.6 to 14.8 Mg ha^{-1} and with a mode of 12.9 Mg ha^{-1} (Figure 3A). For MT, average maize yield was 5.5 Mg ha^{-1}, with IQR ranging from 4.5 to 6.4 Mg ha^{-1} and with a mode of 5.7 Mg ha^{-1} (Figure 3F). In both states, yield frequency distribution was not considered normal according to Shapiro-Wilk test ($P < 0.05$). Furthermore, high within- and between-field variability was documented (Figure 3B and 3G). For RS, field 1 was the most productive with a yield average of 14.9 Mg ha^{-1} with a variation range from 9.2 to 20 Mg ha^{-1} and field 5 was the least productive with a yield average of 10.3 Mg ha^{-1} and ranging from 5 to 15.2 Mg ha^{-1}. For MT, field 10 was the most productive field with a yield average of 7 Mg ha^{-1} with a variation range from 4.1 to 8.6 Mg ha^{-1}, while field 9 was the least productive with a yield average of 4.2 Mg ha^{-1} and with values ranging from 2.9 to 5.9 Mg ha^{-1}.

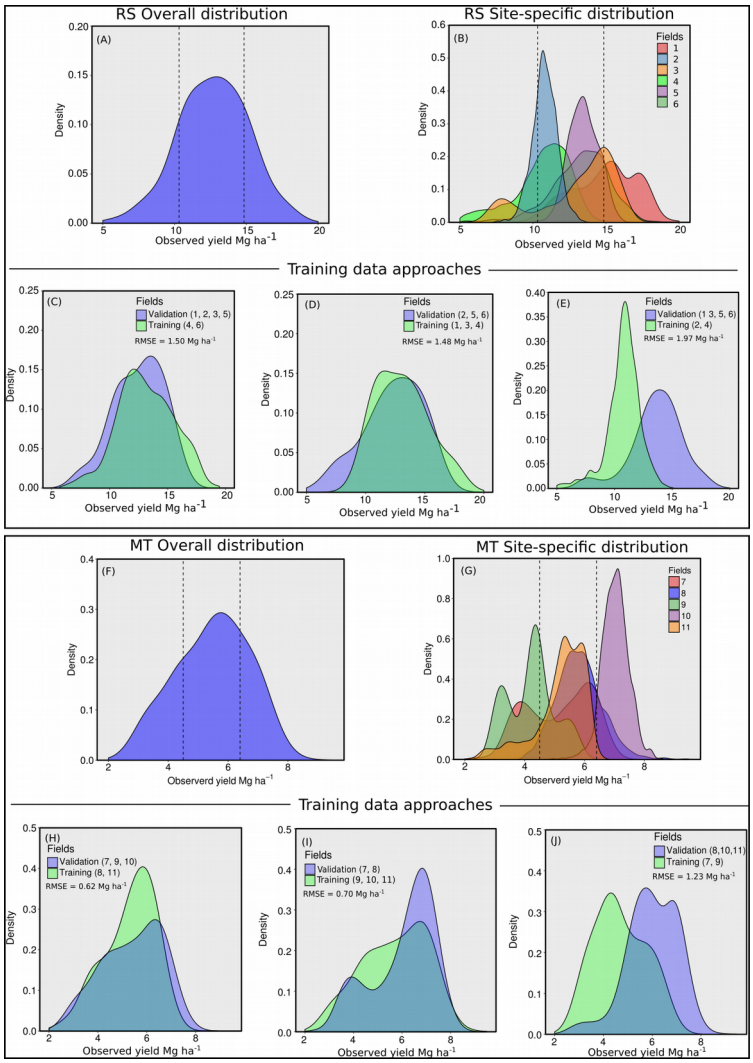


Figure 3. Maize yield frequency distributions for RS (A-E) and MT (F-G) fields. (A and F) Overall yield frequency distribution in RS and MT respectively. (B and G) Field level yield frequency distribution in RS and MT respectively. (C, D, E) Training and validation yield frequency distribution for different training data selection strategies in RS. (H, I, J) Training and validation yield frequency distribution for different training data selection strategies in MT. Root mean square error (RMSE) reported is from the observed and predicted yields using each set of training and validation data.

The field selection to comprise the training data affected the model quality and, consequently, the predictability power of the model. For RS, three different sets of fields were tested as training data: fields 4 and 6 (Figure 3C), fields 1, 3 and 4 (Figure 3D) and fields 2 and 4 (Figure 3E) (Supplementary table 3). The RMSE tended to decrease as the yield frequency distribution of the training data becomes similar to the validation data. When fields

2 and 4 comprised the training data different modes; first and third quartiles ($P < 0.05$) were documented to training and validation data (Supplementary table 4). The aforementioned combination of fields resulted in the highest RMSE (1.97 Mg ha^{-1}). When fields with high amplitude and intermediate yield (compared with all fields) were selected, training and validation data were more alike sharing comparable IQR ($P > 0.05$) with a slightly different mode (12.4 vs. 13 Mg ha^{-1}) ($P < 0.05$) and RMSE of 1.5 Mg ha^{-1}. The lowest RMSE, 1.48 Mg ha^{-1}, was reported when the number of fields increased from 2 to 3, by means the selection of the lowest, the intermediate and the greatest productive field. After this process, training and validation yield frequency distribution resulted in comparable mode ($P > 0.05$) and IQR ($P > 0.05$). Since the selection of one additional field just increased slightly the RMSE (from 1.48 to 1.50 Mg ha^{-1}), only fields 4 and 6 were chosen for posterior analysis, leaving one more field available for the model validation. Likewise, the same criteria aforementioned was applied to MT fields, the selection of the left shifted fields resulted in significantly different modes, first and third quartiles ($P < 0.05$) and the highest RMSE (1.23 Mg ha^{-1}) (Figure 3J). For MT the increase in number of selected fields for comprising training data did not result in the lowest RMSE (0.7 Mg ha^{-1}). This strategy leaded to statistically equal modes ($P >0.05$), but different first and third quartile positions ($P < 0.05$) (Figure 3I). The selection of fields 8 and 11 resulted in non-differences between training and validation modes and IQR ($P > 0.05$) obtaining a RMSE of 0.62 Mg ha^{-1} (Figure 3H). Following the rationale for field selection for RS, the fields 8 and 11 were chosen for posterior analysis since this data training presented the lowest RMSE (Figure 3H) relative to the other tested models (Figure 3I, J). No pattern was observed for skewness and kurtosis linking yield data distribution similarities and RMSE for the models from RS and MT (Supplementary table 4).

3.2. Building forecasting yield models

Spatial autocorrelation analysis conducted using Moran's I test (MI) on VIs and yield data are presented in Supplementary Table 2. In general, autocorrelation (Moran's I test) for all variables was positive and statistically significant (exception for F8) indicating that when yield or VI values are geographically in shorter distances are more alike, diminishing the spatial correlation as the distance increases. The absence of spatial correlation in F8 was probably due to higher yield homogeneity in this field compared to the other ones.

Following the same rationale, forecasting yield models increase predictability power when a spatial correlation structure was considered. The spatial regression models outperformed the OLS once the AIC values were smaller for the spatial models compared to

the OLS ones (Table 2). It indicated that there was a good trade-off between the goodness of fit and the complexity of the model. For the RS model, residuals were assumed following a Gaussian spatial correlation structure, while for the MT and the universal (both RS+MT) models the exponential correlation structure presented the best fit to describe the data (Table 2).

Table 2. Multiple linear regression models for the ordinary least-square (OLS) and regression considering spatial correlation including the vegetation indices (VIs) obtained from mid-season satellite imagery as predictors of the end-season yield monitor data. Equations are related to model with the lowest AIC. SRE = spatial regression considering exponential correlation of the plotted errors. SRG = spatial regression considering Gaussian correlation of the plotted errors. SRS = spatial regression considering spherical correlation of the plotted errors.

Data	Model	AIC	Equation
RS	OLS	3771	Yield (Mg ha^{-1}) =2.7*** + 69.88*** (NDVIre) (R^2 = 0.68)
	SRE	3762	
	SRG	3759	
	SRS	3770	
MT	OLS	3087	Yield (Mg ha^{-1}) = 15.3*** + 81.6*** (NDVIre) – 8.8*** (NDVIG) – 20.3 (NDVI)*** (R^2 = 0.59)
	SRE	891	
	SRG	1986	
	SRS	894	
Universal	OLS	9985	Yield (Mg ha^{-1}) = -25.6*** -46.5 (NDVIre)*** +145.1 (NDVIG)*** – 67.5 (NDVI)*** (R^2 = 0.32)
	SRE	6750	
	SRG	8959	
	SRS	6836	

Notes: The statistically significant coefficients are indicated by asterisks, where * indicates $P < 0.05$; ** indicates $P < 0.01$; and *** indicates $P < 0.001$. Parameters with no asterisks are therefore not significant at the 0.05 level.

All VIs were kept into the MT and universal models after the stepwise selection, while for RS only the NDVIre was retained (Table 2). The NDVIre presented the greatest weight for all models. Even some degree of multicollinearity among the indices was expected since NIR

band was a component of all of them, the VIFs were less than 2 for the VI that remain in the model.

3.3. Spatial and temporal validation of models

In the first step of the spatial validation, the universal model was compared to the site-specific models (state-scale models). The predictability power of the universal model was drastically reduced both for within- (data not shown) and between-field variability (Figure 4A) compared to the site-specific models (Figure 4B). The universal model slightly overestimated yield for the MT fields (low productivity fields) and underestimated yield in RS (high productive fields). Site-specific models resulted in RMSE of 1.5 Mg ha^{-1} for RS and 0.62 Mg ha^{-1} for MT.

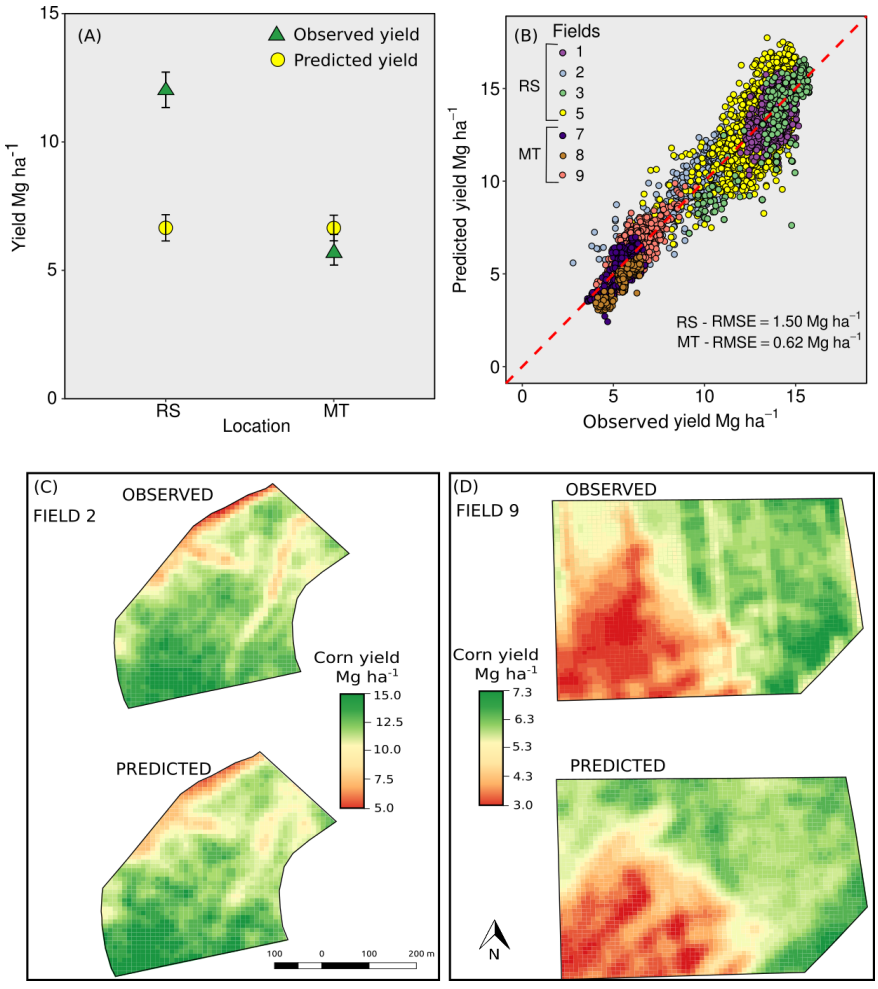


Figure 4. Estimated versus observed maize yield. (A) State-level yield prediction using the Universal forecasting yield model. (B) Within-field yield variability prediction using site-specific maize yield forecasting models. A red dashed line is presented in panel portraying the

1:1 line for the estimated–observed relationship. (C) Observed yield map versus predicted yield map generated based on a site-specific model for RS. (D) Observed yield map versus predicted yield map generated based on a site-specific model for MT. RMSE = Root-mean square error. RS = Rio Grande do Sul. MT = Mato Grosso.

In the second step of the spatial validation, the RS model was used to forecast yield of one additional dataset comprised of six fields located in KS (US). The RS model was chosen for this purpose since the yield frequency distribution of the RS training data was the closest to the one for KS fields (Figure 5A). Despite the similarity in yield frequency distribution for KS and RS, differences in mode and IQR (P > 0.05) were documented. The RS model presented a good predictability in low productive areas and tended to overestimate yield in high productive zones, resulting in a RMSE of 2.22 Mg ha^{-1} (Figure 5B).

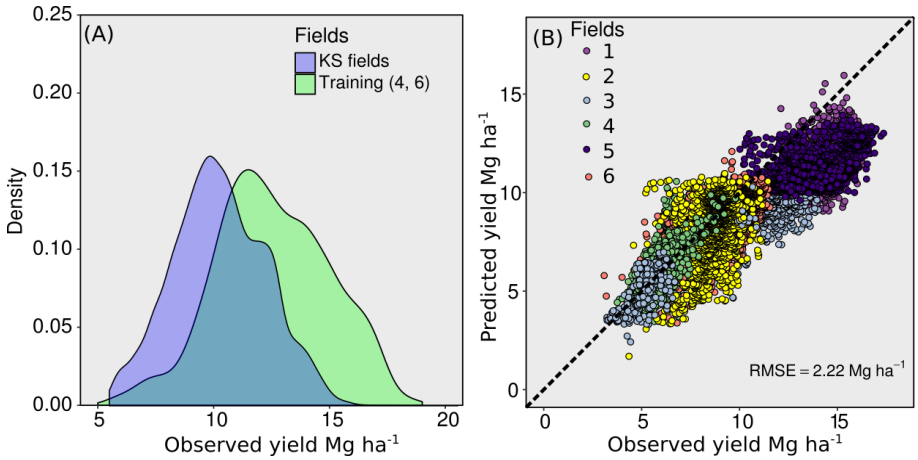


Figure 5. (A) Yield frequency distribution for RS (training data – Fields 4 and 6) and for KS fields and (B) Predicted (estimated via RS yield forecasting model) versus KS observed maize yield (end-season yield monitor data). A dashed black line portrays the 1:1 line for the predicted–observed yield relationship. RMSE = Root-mean square error. RS = Rio Grande do Sul. KS = Kansas.

For the temporal validation, the MT model built with the 2016 data was used to forecast yield for independent fields harvested in 2017. Yield distribution frequency between MT training data (2016) and MT yield data from 2017 was similar with statistically equal mode and IQR (P > 0.05) (Figure 6A). The MT model presented a good predictability power predicting within-field variability of 2017 fields, with a RMSE of 0.95 Mg ha^{-1} (Figure 6B). Historical weather data showed that the 2016 and 2017 growing seasons were similar, with

temperatures slightly above and total precipitation slightly below the average of the last 17 years (period from 1st January to 31st June) (Figure 6C).

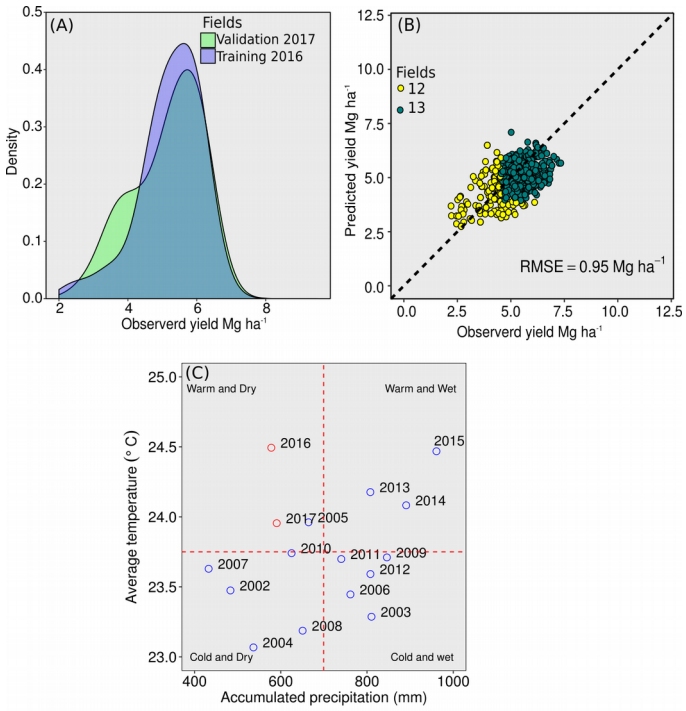


Figure 6. (A) Yield frequency distribution for RS (training data – Fields 4 and 6) and for KS (B) Estimated (predicted via RS yield forecasting model) versus KS observed maize yield (end-season yield monitor data). A dashed black line is presented in panel portraying the 1:1 line for the estimated–observed relationship. (C) Average temperature and accumulated precipitation from last 17 years (period from 1st January to 31st June). A dashed red line represents the average from the entire period. RMSE = Root-mean square error. RS = Rio Grande do Sul. MT = Mato Grosso. KS = Kansas.

4. Discussion

4.1. Building forecasting yield models

Processes to build empirical models usually involve two steps; construction and validation (Becker-Reshef et al., 2010; Hatfield et al., 2008; Peralta et al., 2016). The selection of training and validation data is usually done randomly (Assefa et al., 2016; Lopresti et al., 2015; Peralta et al., 2016), but it is predictable that the selected training data can affect directly the model predictability power (Schwalbert et al., 2018; Sheridan, 2013). The first outcome of this study was a relationship between similarity of training and validation

data and predictability power of the model. Statistics parameters such as mode, first and third quartiles, were implemented to test the similarity between datasets. The selection of fields with a high variability and not shifted (related to the overall yield frequency distribution) increased the likelihood of obtaining more representative models. Fields with a high degree of uniformity in yield are not expected to add useful information to the model related to the yield-VI relationship (Peralta et al., 2016). Fields with left or right shift on the yield frequency distributions related to the overall yield frequency distribution (when all fields were aggregated) also can bias the model. Left shifted fields (with yields towards low values) could have a yield-VI relationship affected by biotic or abiotic stress condition after image acquisition (Sadras and Calviño, 2001), while right shifted fields (with yields towards high values) could face problems related to saturation of VIs, such as NDVI (Hatfield et al., 2008). This study tested different statistical parameters (mean, mode, first and third quantile positions, skewness and kurtosis) as potential indicators of similarities in yield frequency data distribution providing guidelines for selection of ground truth data for build in season forecast models. Mode and the quantile positions were most suitable parameters driving the selection for the set of training and validation data that minimize the RMSE (Supplementary table 4). Similar results were reported by Schwalbert et al. (2018) in a study involving maize yield response to plant density and nitrogen rates. In summary, this study also presents a novel approach for the selection of the ground-truth training data utilized for building forecasting yield models based on studying data yield distribution.

Additionally, the approach used to build the yield forecasting models as well as the selection of the VIs influenced model predictability. The approach considering spatial correlation of the regression residuals outperformed the method considering the i.i.d assumption. This result is expected, since the positive spatial correlation for yield data and for VI is already well-known (Bakhsh et al., 2000; Bressler et al., 1981,1982; Jaynes and Colvin, 1997; Peralta et al., 2016; Morkoc et al., 1985, Timlin et al., 1998) and therefore spatial correlation of regression residuals should be accounted for (Anselin et al., 2004; DiRienzo et al., 2000; Leiser et al., 2012; Peralta et al., 2016). Despite that, still there is a few number of studies showing the benefits of spatial adjustment to models predicting yield from imagery data (Imran etl al., 2013; Peralta et al., 2016). Regarding the performance of the VIs as explanatory variables, NDVIre presented the highest weight in the regression and it was also the most retained index. Recently, Peralta et al. (2016) also reported that this VI was more effective to predicted yields relative to NDVI and NDVIG. The explanation for this is that the NDVIre is less influenced by changes in leaf area avoiding saturation issues at medium to

high LAI and yield. It is imperative also mentioned that for the MT and universal models, NDVIG and NDVI were also retained, reflecting the potential of these indices for predicting yield variation and to fine-tune the proposed yield forecasting model.

4.2. Spatial and temporal validation of models

Empirical models are frequently reported as an efficient tool to forecast cereal yield, and variations in VI can account for more than 80% of the observed variation in yields within individual fields (Shanahan et al., 2001; Wiegand and Richardson, 1990). Despite the high capacity to explain yield variability, even within-field, empirical models are known to be regionals (Becker-Reshef et al., 2010; Doraiswamy et al., 2003; Hatfield et al., 2008; Moriondo et al., 2007). Similar constraint was documented in this study since the universal model was not suitable even to forecast yield variations in a state-scale. When forecasting yield models were applied individually for MT or RS, the predictability power increased substantially. The overall yields were lower in MT than in RS because the maize in MT was not grown during the best season for that region (second season). The maize in MT was affected adversely by abiotic stresses (Minuzzi et al., 2015) due to the season. When satellite imagery was obtained prior to flowering, further abiotic stress in these fields could severely affect final yield (Sadras and Calviño, 2001) and consequently the yield-VI relationship. Truly, the model is forecasting the potential yield at the flowering, and that is the reason why in some conditions there was an overestimation in the prediction, as visualized in figure 5B, for high yield values. Furthermore, as with any purely empirical approach, extrapolation of equations to new locations or years can be problematic (Hatfield et al., 2008; Lobell, 2013; Lopresti et al., 2015; Moriondo et al., 2007). For this study, the yield frequency distribution of RS and KS fields were quite similar resulting in reasonable yield predictability despite a loss in sensibility to explain within-field yield variability, highlighted by the increase in RMSE in relation to the forecast for the RS fields. Another example that in determined conditions empirical models could overcome the spatial constraint is the study developed by Becker-Reshef et al. (2010), where models developed in KS were successfully applied to forecasting wheat yields in Ukraine. In the same way as the distance in space (geographic distance), distances in time (years) are expected to decrease the predictability of the model (Bognár et al., 2011). However, in our study, weather conditions lead to similar growth environments resulting in comparable yield frequency distributions between 2016 and 2017 seasons (Fig. 6C); therefore, the model predictability was just slightly affected but quite alike. Despite that,

the temporal analysis should be cautiously evaluated since it comprises one year and a specific region around the globe. Further testing including more years and other regions presenting comparable weather conditions should be pursued to validate this point.

This study showed that the selection of the fields for comprising training data affected directly the model structure. Historical yield information is available in platforms such as National Agricultural Statistics Service (NASS), and once knowing the overall yield frequency distribution from a specific region, fields representative to the region can be selected to scale-up the yield forecasting models to county, agricultural districts, and state-scales. One of the main drawbacks of remotely sensed based empirical models for estimating yields has been that their application is valid only for the areas they have been calibrated for (Doraiswamy et al., 2003; Hatfield et al., 2008; Lobell, 2013). By means of the current outcomes presented in this study, it can be inferred that independent datasets could portray in a high-probability comparable yield-VI relationship if the following criteria are fulfilled: yield data distribution with, i) IQR, ii) mode statistically similar, and satellite imagery, iii) collected at a similar growth stage, even with fields separated by space or time. The latter could provide a foundational knowledge to establish conditions (regions in space and year characteristics) where determined empirical models could be suitable, and when a new model should be developed. Furthermore, this study may provide guidelines for applicability of yield forecast models where ground-truth data is limited or scarce, providing fundamental information for supporting policy formulation and helping farmers, consumers, researchers, providing guidelines for making informed decisions based on the crop yield forecast report. Therefore, standards or basis of how to collect data for building more accurately forecasting yield models and information regarding the applicability of those models are extremely important and useful.

5. Conclusions

The likelihood of two independent datasets portray comparable yield-IV relationship increases as their yield data distribution becomes more alike, mainly related to the position of the mode, first and third quartiles (IQR). In this current study model performance was more affected by differences in the yield frequency distribution rather than by distance in space (BR and KS) or time (2016 and 2017 seasons). Since RS and MT presented a large difference in yield frequency distribution, the universal model to estimate maize yield in both states

presented small predictability power compared to the site-specific models (individual model per state).

The regression model using the NDVI, NDVIG, and NDVIre showed high performance for predicting within-field yield variability. Approaches that adequately account for spatial correlation outperformed the OLS models since yield and VIs were spatially correlated

This current analysis is among the few studies demonstrating the utilization of mid-season high-resolution satellite imagery to forecasting within-field maize yield variation. Future research should be focused on improving the understanding of historical yield distribution at larger scales (county, district or state-level) aiming at mapping the potential and limitations of scaling-up yield forecasting models.

Acknowledgements: This study was supported by CAPES Foundation, Ministry of Education of Brazil, Brasilia - DF, Zip Code 70.040-020, Aquarius project (http://w3.ufsm.br/projetoaquarius/index.php/pt/), and Kansas Corn Commission. This is contribution no. 18-073-J from the Kansas Agricultural Experiment Station and process 88887.130848/2016-00 from CAPES.

References

Alganci, U., Ozdogan, M., Sertel, E., & Ormeci, C., (2014). Estimating maize and cotton yield in southeastern Turkey with integrated use of satellite images, meteorological data and digital photographs. F. Crop. Res. 157, 8–19. doi:10.1016/j.fcr.2013.12.006

Amidan, B.G., Ferryman, T.A., & Cooley, S.K., (2005). Data outlier detection using the chebyshev theorem, in: IEEE Aerospace Conference Proceedings. doi:10.1109/AERO.2005.1559688

Anselin, L., (1995). Local Indicators of Spatial Association—LISA. Geogr. Anal. 27, 93–115. doi:10.1111/j.1538-4632.1995.tb00338.x

Anselin, L., Bongiovanni, R., & Lowenberg-deboer, J., (2004). A Spatial Econometric Approach To the Economics of Site Specific Nitrogen Management in Corn Production. Am. J. Agric. Econ. 86, 675–687.

Assefa, Y., Prasad, P.V.V., Carter, P., Hinds, M., Bhalla, G., & Ciampitti, I.A., 2016. Yield Responses to Planting Density for US Modern Corn Hybrids : A Synthesis-Analysis. Crop Sci. 56, 1–38. doi:10.2135/cropsci2016.04.0215

Azzari, G., Jain, M., & Lobell, D. B. (2016). Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries. Remote Sensing of Environment, 129–141. https://doi.org/10.1016/j.rse.2017.04.014

Bakhsh, A., Jaynes, D. B., Colvin, T. S., & Kanwar, R. S., (2000). Spatio-temporal analysis of yield variability for a corn-soybean field in Iowa. Trans. ASAE 43, 31–38.

Ban, H. Y., Kim, K. S., Park, N. W., & Lee, B. W., (2017). Using MODIS data to predict regional corn yields. Remote Sens. 9. doi:10.3390/rs9010016

Becker-Reshef, I., Vermote, E., Lindeman, M., & Justice, C., (2010). A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data. Remote Sens. Environ. 114, 1312–1323. doi:10.1016/j.rse.2010.01.010

Bivand, R., Piras, G., (2015). Comparing Implementations of Estimation Methods for Spatial Econometrics. J. Stat. Softw. 63, 1–36.

Bognár, P., Ferencz, C., Pásztor, S., Molnár, G., Timár, G., Hamar, D., Lichtenberger, J., Székely, B., Steinbach, P., & Ferencz, O. E., (2011). Yield forecasting for wheat and corn in Hungary by satellite remote sensing. Int. J. Remote Sens. 32, 4759–4767. doi:10.1080/01431161.2010.493566

Bongiovanni, R. G., Robledo, C. W., & Lambert, D. M., (2007). Economics of site-specific nitrogen management for protein content in wheat. Comput. Electron. Agric. 58, 13–24. doi:10.1016/j.compag.2007.01.018

Bu, H., Sharma, L. K., Denton, A., & Franzen, D. W., (2017). Comparison of satellite imagery and ground-based active optical sensors as yield predictors in sugar beet, spring wheat, corn, and sunflower. Agron. J. 109, 299–308. doi:10.2134/agronj2016.03.0150

Canty, A., & Ripley, B., (2017). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-19.

Congedo, L., (2016). Semi-Automatic Classification Plugin - User Manual. doi:http://dx.doi.org/10.13140/RG.2.1.1219.3524

Córdoba, M. A., Bruno, C.I., Costa, J. L., Peralta, N. R., & Balzarini, M. G., (2016). Protocol for multivariate homogeneous zone delineation in precision agriculture. Biosyst. Eng. 143, 95–107. doi:10.1016/j.biosystemseng.2015.12.008

Deering, D. W., (1978). Rangeland reflectance characteristics measured by aircraft and spacecraft sensors. Ph. D. thesis, Texas A&M Univ., Coll. Stn. 338.

DiRienzo, C., Fackler, P., & Goodwin, B. K., (2000). Modeling spatial dependence and spatial heterogeneity in county yield forecasting models, in: American Agricultural Economics Association Annual Meeting, Tampa, Florida, July.

Doraiswamy, P. C., Moulin, S., Cook, P. W., & Stern, A., (2003). Crop Yield Assessment from Remote Sensing. Photogramm. Eng. Remote Sensing 69, 665–674. doi:10.14358/PERS.69.6.665

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., … Bargellini, P. (2012). Sentinel-2: ESA’s Optical High-Resolution Mission for GMES Operational

Services. *Remote Sensing of Environment*, *120*, 25–36. https://doi.org/10.1016/j.rse.2011.11.026

Efron, B., & Tibshirani, R. J., (1993). An introduction to the bootstrap. Refrig. Air Cond. 57, 436. doi:10.1111/1467-9639.00050

Gitelson, A. A., Kaufman, Y. J., & Merzlyak, M. N., (1996). Use of a green channel in remote sensing of global vegetation from EOS-MODIS. Remote Sens. Environ. 58, 289–298. doi:10.1016/S0034-4257(96)00072-7

Gitelson, A. A., & Merzlyak, M. N., (1994). Spectral Reflectance Changes Associated with Autumn Senescence of Aesculus-hippocastanum L. and Acer-platanoides L. Leaves - Spectral Features and Relation to Chlorophyll Estimation. J. Plant Physiol. 143, 286–292. doi:10.1016/S0176-1617(11)81633-0

Gholap, J., A. Ingole, J. Gohil, S. Gargade, & V. Attar. (2012). Soil data analysis using classification techniques and soil attribute prediction. International Journal of Computer Science Issues 9(3):415–418.

Gonzalez-Sanchez, A, (2014). Predictive ability of machine learning methods for massive crop yield prediction. Spanish J. Agric. Res. 12, 313–328. doi:10.5424/sjar/2014122-4439

Haboudane, D., Miller, J. R., Pattey, E., Zarco-Tejada, P. J., & Strachan, I. B., (2004). Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture. Remote Sens. Environ. 90, 337–352. doi:10.1016/j.rse.2003.12.013

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L., (1998). Multivariate data analysis. Prentice hall Upper Saddle River, NJ.

Hamada, Y., Ssegane, H., & Negri, M. C., (2015). Mapping intra-field yield variation using high resolution satellite imagery to integrate bioenergy and environmental stewardship in an agricultural watershed. Remote Sens. 7, 9753–9768. doi:10.3390/rs70809753

Hamar, D., Ferencz, C., Lichtenberger, J., Tarcsai, G., & Ferencz-Arkos, I., (1996). Yield estimation for corn and wheat in the Hungarian Great Plain using Landsat MSS data. Int. J. Remote Sens. 17, 1689–1699. doi:10.1080/01431169608948732

Hatfield, J. L., Gitelson, A. A., Schepers, J. S., & Walthall, C. L., (2008). Application of spectral remote sensing for agronomic decisions. Agron. J. doi:10.2134/agronj2006.0370c

Horbe, T. A. N., Amado, T. J. C., Reimche, G. B., Schwalbert, R. A., Santi, A. L., & Nienow, C., (2016). Optimization of within-row plant spacing increases nutritional status and corn yield: A comparative study. Agron. J. 108, 1962–1971. doi:10.2134/agronj2016.03.0156

Imran, M., Zurita-Milla, R., & Stein, A., (2013). Modeling crop yield in West - African rainfed agriculture using global and local spatial regression. Agron. J. 105, 1177–1188.

Jaynes, D. B., & Colvin, T. S., (1997). Spatiotemporal Variability of Corn and Soybean Yield. Agron J 89, 30–37.

Jin, Z., Azzari, G., Burke, M., Aston, S., & Lobell, D. (2017). Mapping Smallholder Yield Heterogeneity at Multiple Scales in Eastern Africa. *Remote Sensing*, *9*(9), 931. https://doi.org/10.3390/rs9090931

Johnson, D. M., (2014). An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. Remote Sens. Environ. 141, 116–128. doi:10.1016/j.rse.2013.10.027

Johnson, J. B., & Omland, K. S., (2004). Model selection in ecology and evolution. Trends Ecol. Evol. 19, 101–108. doi:10.1016/j.tree.2003.10.013

Kantanantha, N., Serban, N., & Griffin, P., (2010). Yield and price forecasting for stochastic crop decision planning. J. Agric. Biol. Environ. Stat. 15, 362–380. doi:10.1007/s13253-010-0025-7

Leiser, W. L., Rattunde, H. F., Piepho, H. P., & Parzies, H. K., (2012). Getting the Most Out of Sorghum Low-Input Field Trials in West Africa Using Spatial Adjustment. J. Agron. Crop Sci. 198, 349–359. doi:10.1111/j.1439-037X.2012.00529.x

Lobell, D. B., (2013). The use of satellite data for crop yield gap analysis. F. Crop. Res. 143, 56–64. doi:10.1016/j.fcr.2012.08.008

Lobell, D. B., Thau, D., Seifert, C., Engle, E., & Little, B., (2015). A scalable satellite-based crop yield mapper. Remote Sens. Environ. 164, 324–333. doi:10.1016/j.rse.2015.04.021

Lopresti, M. F., Di Bella, C. M., & Degioanni, A. J., (2015). Relationship between MODIS-NDVI data and wheat yield: A case study in Northern Buenos Aires province, Argentina. Inf. Process. Agric. 2, 73–84. doi:10.1016/j.inpa.2015.06.001

Minuzzi, R. B., Lopes, F. Z., Minuzzi, R. B., & Lopes, F. Z., (2015). Desempenho agronômico do milho em diferentes cenários climáticos no Centro-Oeste do Brasil. Rev. Bras. Eng. Agrícola e Ambient. 19, 734–740. doi:10.1590/1807-1929/agriambi.v19n8p734-740

Monteith, J.L., (1977). Climate and the efficiency of crop production in Britain. Philos. Trans. R. Soc. Lond. B 281, 277–294

Moriondo, M., Maselli, F., & Bindi, M., (2007). A simple model of regional wheat yield based on NDVI data. Eur. J. Agron. 26, 266–274. doi:10.1016/j.eja.2006.10.007

Mulla, D.J., (2013). Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. Biosyst. Eng. 114, 358–371. doi:10.1016/j.biosystemseng.2012.08.009

Myneni, R. B., & D. L. Williams. "On the relationship between FAPAR and NDVI." Remote Sensing of Environment 49.3 (1994): 200-211.

Nguy-Robertson, A., Gitelson, A., Peng, Y., Viña, A., Arkebauer, T., & Rundquist, D., (2012). Green leaf area index estimation in maize and soybean: Combining vegetation indices to achieve maximal sensitivity. Agron. J. 104, 1336–1347. doi:10.2134/agronj2012.0065

Noureldin, N.A., Aboelghar, M.A., Saudy, H.S., & Ali, A.M., (2013). Rice yield forecasting models using satellite imagery in Egypt. Egypt. J. Remote Sens. Sp. Sci. 16, 125–131. doi:10.1016/j.ejrs.2013.04.005

Pebesma, E.J., (2004). Multivariable geostatistics in S: the gstat package. Comput. Geosci.

Peralta, N., Assefa, Y., Du, J., Barden, C., & Ciampitti, I., (2016). Mid-Season High-Resolution Satellite Imagery for Forecasting Site-Specific Corn Yield. Remote Sens. 8, 848. doi:10.3390/rs8100848

Peralta, N.R., & Costa, J.L., (2013). Delineation of management zones with soil apparent electrical conductivity to improve nutrient management. Comput. Electron. Agric. 99, 218–226. doi:10.1016/j.compag.2013.09.014

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team, (2017). nlme: Linear and Nonlinear Mixed Effects Models.

Raun, W.R., Solie, J.B., & Johnson, G.V., (2002). Improving nitrogen use efficiency in cereal grain production with optical sensing and variable rate application. Agron. J. 94, 815–820.

Reeves, M. C., Zhao, M., & Running, S. W., (2005). Usefulness and limits on MODIS GPP for estimating wheat yield. Int. J. Remote Sens. 26, 1403–1421. doi:10.1080/01431160512331326567

Rembold, F., Atzberger, C., Savin, I., & Rojas, O., (2013). Using low resolution satellite imagery for yield prediction and yield anomaly detection. Remote Sens. doi:10.3390/rs5041704

Ribeiro Jr, P. J., & Diggle, P.J., (2016). geoR: Analysis of Geostatistical Data.

Rouse, J., Haas, R., & Schell, J., (1974). Monitoring the vernal advancement and retrogradation (greenwave effect) of natural vegetation. Texas A M Univ. 1–8.

Sadras, V. O., & Calviño, P. A., (2001). Quantification of grain yield response to soil depth in soybean, maize, sunflower, and wheat. Agron. J. 93, 577–583. doi:10.2134/agronj2001.933577x

Sakamoto, T., Gitelson, A. A., & Arkebauer, T. J. (2014). Near real-time prediction of U.S. corn yields based on time-series MODIS data. *Remote Sensing of Environment*, *147*, 219–231. https://doi.org/10.1016/j.rse.2014.03.008

Schueller, J. K., & Bae, Y. H., (1987). Spatially attributed automatic combine data acquisition. Comput. Electron. Agric. 2, 119–127. doi:10.1016/0168-1699(87)90022-6

Schwalbert, R., Amado, T., Tiago, H., Lincon, S., Yared, A., Prasad, V., … Ciampitti, I. (2018). Corn yield response to plant density and nitrogen: Spatial model and yield distribution. *Agron. J. (First Look)*.

Shanahan, J. F., Schepers, J. S., Francis, D. D., Varvel, G. E., Wilhelm, W. W., Tringe, J. M., Schlemmer, M. R., & Major, D. J., (2001). Use of Remote-Sensing Imagery to Estimate Corn Grain Yield. Agron. J. 93, 583–589. doi:10.2134/agronj2001.933583x

Shao, Y., Campbell, J. B., Taff, G. N., & Zheng, B., (2015). An analysis of cropland mask choice and ancillary data for annual corn yield forecasting using MODIS data. Int. J. Appl. Earth Obs. Geoinf. 38, 78–87. doi:10.1016/j.jag.2014.12.017

Sheridan, R.P., (2013). Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. J. Chem. Inf. Model. 53, 783–790. doi:10.1021/ci400084k

Sibley, A. M., Grassini, P., Thomas, N. E., Cassman, K. G., & Lobell, D. B., (2014). Testing remote sensing approaches for assessing yield variability among maize fields. Agron. J. 106, 24–32. doi:10.2134/agronj2013.0314

Solie, J. B., Dean Monroe, A., Raun, W. R., & Stone, M. L., (2012). Generalized algorithm for variable-rate nitrogen application in cereal grains. Agron. J. 104, 378–387. doi:10.2134/agronj2011.0249

Stafford, J. V., Ambler, B., Lark, R. M., & Catt, J., (1996). Mapping and interpreting the yield variation in cereal crops. Comput. Electron. Agric. 14, 101–119. doi:10.1016/0168-1699(95)00042-9

Stone, R. C., & Meinke, H., (2005). Operational seasonal forecasting of crop performance. Philos. Trans. R. Soc. B Biol. Sci. 360, 2109–2124. doi:10.1098/rstb.2005.1753

Timlin, D., Pachepsky, Y., Snyder, V., & Bryant, R. B., (1998). Spatial and temporal variability of corn grain yield on a hillslope. Soil Sci. Soc. Am. J. 62, 764–773.

Tucker, C. J., (1979). Red and Photographic Infrared linear Combinations for Monitoring Vegetation. Remote Sens. Environ. 8, 127–150.

Venables, W. N., & Ripley, B. D., (2002). Modern Applied Statistics with S, Fourth. ed. Springer, New York.

Wall, L., Larocque, D., & Léger, P.-M., (2008). The early explanatory power of NDVI in crop yield modelling. Int. J. Remote Sens. 29, 2211–2225. doi:10.1080/01431160701395252

Wiegand, C., & Richardson, A., (1990). Use of spectral vegetation indices to infer leaf area, evapotranspiration and yield: I. rationale. Agron. J. 82, 623–629. doi:10.2134/agronj1990.00021962008200030038x

Zuur, A. F., Ieno, E. N., & Elphick, C. S., (2010). A protocol for data exploration to avoid common statistical problems. Methods Ecol. Evol. 1, 3–14. doi:10.1111/j.2041-210X.2009.00001.x

5 DISCUSSÃO

Embora imagens de satélites ainda tenham um papel limitado na maioria dos esforços operacionais para monitorar a produtividade, vários estudos recentes permitiram o progresso em direção ao uso mais rotineiro dessa ferramenta, impulsionados pelo acesso facilitado às imagens nos últimos anos devido ao maior número de satélites em órbita, e pelas novas tecnologias de processamento em nuvem como o GEE, permitindo a manipulação, armazenamento e processamento de grande volume de dados, possibilitando o desenvolvimento de novos algoritmos mais generalizáveis (AZZARI et al., 2017).

Apesar da evidente evolução nas técnicas de mapeamento de produtividade, sua aplicabilidade regional ainda é limitada pela dificuldade de obtenção de informações confiáveis sobre a distribuição geográfica das áreas agrícolas (JIN et al., 2017; SAKAMOTO; GITELSON; ARKEBAUER, 2014; SHELESTOV et al., 2017). As primeiras contribuições desse estudo foram: a) a proposição de um modelo capaz de mapear a localização espacial das áreas produtoras de milho no *Corn Belt* americano com uma acurácia superior a 80% e predizer a produtividade dessa cultura com um erro médio absoluto inferior a 0,9 Mg ha^{-1}, com uma antecedência de 98 dias em relação a colheita. Esse modelo foi amplamente validado para diferentes condições de clima e solo e pode ser aplicado para diferentes regiões produtoras que possuam um nível de informação semelhante ao encontrado na região onde o modelo foi proposto; e b) a proposição de um modelo para predizer a produtividade da cultura da soja no estado do Rio Grande do Sul sem o uso de uma camada de informação específica de cultura, capaz de predizer a produtividade dessa cultura com um erro absoluto médio de 0,24 Mg ha^{-1} com uma antecedência de aproximadamente 40 dias em relação à colheita. Os dois modelos foram testados para diferentes datas, e apesar de apresentarem um erro crescente à medida que a data da previsão é antecipada eles ainda apresentaram um desempenho satisfatório para previsões 70 dias antes da colheita para a cultura da soja com um erro médio absoluto de 0,42 Mg ha^{-1}, e 122 dias antes da colheita do milho com um erro médio absoluto inferior a 1 Mg ha^{-1}.

Uma segunda contribuição desse estudo está relacionado à incorporação dos dados climáticos no modelo preditivo. Apesar dos índices de vegetação indiretamente captarem o efeito das variáveis meteorológicas sobre o desenvolvimento vegetal, uma substancial melhora no desempenho dos modelos foi documentada quando temperatura, precipitação e DPV foram incluídas no modelo juntamente aos índices de vegetação. Apesar desse efeito já

ter sido reportado anteriormente na literatura (PENG et al., 2018) apenas uma pequena fração dos modelos propostos faz uso dessa fonte de informação adicional.

Não obstante, considerável melhora no desempenho dos modelos de previsão foi documentada pelo uso de redes neurais de aprendizagem profunda. Essa técnica representa uma extensão das redes neurais convencionais apresentando diversas camadas de abstração o que possibilita o modelo representar complexas interações entre as variáveis explanatórias e a variável resposta (CUNHA; SILVA; NETTO, 2018; KHAKI; WANG, 2019; YOU et al., 2017). Nesse trabalho em especial, foram usadas uma classe específica de redes neurais, conhecidas como *Long Short Term Memory – LSTM*. Redes neurais dessa natureza são adequadas para reconhecer padrões em séries temporais (como os dados provenientes das imagens de satélite e dados meteorológicos) e normalmente apresentam desempenho superior com dados dessa natureza (YOU et al., 2017).

Com exceção ao desafio de identificar a localização espacial das áreas agrícolas, estabelecer relações empíricas entre produtividade e preditores normalmente é uma tarefa menos complexa para domínios maiores, como municípios, condados, estados, etc., comparados com pequenas áreas produtivas. As dificuldades de estabelecer essas relações matemáticas em domínios menores (maior resolução) está relacionada a aspectos como a menor disponibilidade de séries históricas de satélites com alta resolução espacial e temporal e a maior dificuldade de coletar dados de produtividade em quantidade suficiente para treinar e validar modelos adequadamente. Além disso, a capacidade de generalização (aplicação do modelo além das condições onde ele foi parametrizado) permanece uma grande incógnita para relações empíricas estabelecidas localmente. A quarta contribuição desse estudo foi fornecer diretrizes a fim de estabelecer limites para generalizações espaço-temporais de modelos empíricos locais. Os resultados desse estudo demonstram que similaridades na distribuição de frequência dos dados de produtividade usados para treinar os modelos são mais importantes que distâncias geográficas (espaciais) ou temporais (anos). Foi documentado que modelos ajustados para o estado do Rio Grande do Sul tiveram um desempenho superior quando aplicados no estado do Kansas – EUA (distribuições de frequência de produtividade semelhantes) em relação às áreas localizadas no estado do Mato Grosso (segunda safra) (distribuição de frequência não-similares). Em relação à escala temporal, modelos ajustados para o ano agrícola de 2016 tiveram um desempenho satisfatório para a safra 2017, considerando que as duas safras tiveram condições meteorológicas similares o que resultou em patamares de produtividades comparáveis.

A principal contribuição dessa pesquisa é mostrar os benefícios potenciais da integração de técnicas estatísticas, dados de sensoriamento remoto e dados meteorológicos na estimativa de produtividade de culturas agrícolas em diferentes escalas geográficas. Entretanto, vale ressaltar que existe oportunidade para aprimorar os resultados apresentados nesse estudo através do(a): i) uso de imagens de satélites comercias com maior resolução temporal e espacial como Rapid-Eye, Skysat and WorldView, ii) uso de novas tecnologias embarcas em satélites baseadas na fluorescência da clorofila que estarão disponíveis em um futuro próximo (DRUSCH et al., 2017), e iii) fusão de modelos empíricos e mecanísticos para aumentar a capacidade de generalização dos modelos preditivos.

6 CONCLUSÃO

Modelos de preditivos baseados em imagens de satélite e variáveis meteorológicas podem antecipar informações de produtividade da cultura do milho em até 122 dias em relação à data de colheira com um erro menor que 1 Mg ha^{-1}, e em 70 dias para a cultura da soja com um erro de 0,42 Mg ha^{-1} em nível municipal no estado do Rio Grande do Sul – Brasil. Espera-se que o erro associado as previsões diminua a medida que as previsões sejam realizadas em datas mais próximas à colheita. Duas diferentes abordagens foram testadas com sucesso nesse estudo para filtrar *pixels* de interesse das imagens de satélite e remover informações de alvos não desejados: a) o uso de informações de anos anteriores para treinar modelos de classificação usando imagens de satélite, capazes de identificar áreas agrícolas em tempo real. Esses modelos foram capazes de atingir valores de acurácia superiores à 85% para a cultura do milho nos EUA; e b) uso de informações de acesso público e georreferenciadas de áreas agrícolas, porém não específicas para a cultura de interesse. Essa segunda abordagem funcionou para o estado do Rio Grande do Sul, porém deve-se destacar que o sistema de rotação de cultura de verão nessa região têm predominância de duas culturas, soja e milho, com uma frequência de ocorrência muito maior da primeira, que foi a cultura considerada no modelo preditivo.

A incorporação de variáveis meteorológicas nos modelos preditivos se mostra uma abordagem promissora com potencial para aumentar a assertividade das previsões. O uso conjunto de dados de sensoriamento remoto e meteorológicos oferece uma oportunidade de coleta de dados em um volume sem precedentes suficientes para treinar modelos mais complexos baseados em *deep learning* capazes de retratar complexas interações entre a

variável resposta e os preditores com potencial para superar os algoritmos convencionalmente usados.

Modelos preditivos empíricos locais possuem menor capacidade de generalização em decorrência da limitada quantidade de dados nessa escala, tanto da variável resposta como dos preditores. Esse estudo objetivou fornecer diretrizes a fim de determinar limites para extrapolação espaço-temporal de tais modelos. Os resultados apesar de promissores, ainda podem ser considerados incipientes e novas abordagens incluindo o uso de modelos mecanísticos baseados em processos fornece uma ótima oportunidade para geração de pseudo-observações capazes aumentar a capacidade de generalização desses modelos em diversas ordens de magnitude.

REFERÊNCIAS

AZZARI, G.; JAIN, M.; LOBELL, D. B. Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries. **Remote Sensing of Environment**, v. 202, p. 129–141, 2017.

BLOOM, J.; STUART, F.; MOONEY, A. Rsource limitation in plantas - An economic analogy. **Annual review of Ecology and Systematics**, v. 16, n. 1, p. 363–392, 1985.

CLEVERS, J. A simplified approach for yield prediction of sugar beet based on optical remote sensing data. **Remote Sensing of Environment**, v. 61, n. 2, p. 221–228, 1997.

CUNHA, R. L. F.; SILVA, B.; NETTO, M. A. S. A scalable machine learning system for pre-season agriculture yield forecast. **Proceedings - IEEE 14th International Conference on eScience, e-Science 2018**, p. 423–430, 2018.

DENTE, L. et al. Assimilation of leaf area index derived from ASAR and MERIS data into CERES-Wheat model to map wheat yield. **Remote Sensing of Environment**, v. 112, n. 4, p. 1395–1407, 2008.

DIRIENZO, C.; FACKLER, P.; GOODWIN, B. K. **Modeling spatial dependence and spatial heterogeneity in county yield forecasting models**. American Agricultural Economics Association Annual Meeting, Tampa, Florida, July. **Anais**...2000Disponível em: <http://ageconsearch.umn.edu/bitstream/21763/1/sp00di01.pdf>. Acesso em: 22 jul. 2017

DORAISWAMY, P. C. et al. Application of MODIS derived parameters for regional crop yield assessment. **Remote Sensing of Environment**, v. 97, n. 2, p. 192–202, 2005.

DRUSCH, M. et al. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. **Remote Sensing of Environment**, v. 120, p. 25–36, 2012.

DRUSCH, M. et al. The FLuorescence EXplorer Mission Concept—ESA's Earth Explorer 8. **IEEE Transactions on Geoscience and Remote Sensing**, v. 55, n. 3, p. 1273–1284, 2017.

GALLEGO, J.; CARFAGNA, E.; BARUTH, B. Accuracy, Objectivity and Efficiency of Remote Sensing for Agricultural Statistics. **Agricultural Survey Methods**, p. 193–211, 2010.

GORELICK, N. et al. Google Earth Engine: Planetary-scale geospatial analysis for everyone. **Remote Sensing of Environment**, v. 202, p. 18–27, 2017.

JIN, Z. et al. Mapping Smallholder Yield Heterogeneity at Multiple Scales in Eastern Africa. **Remote Sensing**, v. 9, n. 9, p. 931, 2017.

JIN, Z.; AZZARI, G.; LOBELL, D. B. Improving the accuracy of satellite-based high-resolution yield estimation: A test of multiple scalable approaches. **Agricultural and Forest Meteorology**, v. 247, p. 207–220, 2017.

JOHNSON, D. M. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. **Remote Sensing of Environment**, v. 141, p. 116–128, fev. 2014.

KHAKI, S.; WANG, L. Crop Yield Prediction Using Deep Neural Networks. 2019.

LAUNAY, M.; GUERIF, M. Assimilating remote sensing data into a crop model to improve predictive performance for spatial applications. **Agriculture, Ecosystems and Environment**, v. 111, n. 1–4, p. 321–339, 2005.

LOBELL, D. B. The use of satellite data for crop yield gap analysis. **Field Crops Research**, v. 143, p. 56–64, 2013.

LOBELL, D. B. et al. Greater sensitivity to drought accompanies maize yield increase in the U.S. Midwest. **Science**, v. 344, n. 6183, p. 516–519, 2014.

LOBELL, D. B. et al. A scalable satellite-based crop yield mapper. **Remote Sensing of Environment**, v. 164, p. 324–333, 2015.

LOPRESTI, M. F.; DI BELLA, C. M.; DEGIOANNI, A. J. Relationship between MODIS-NDVI data and wheat yield: A case study in Northern Buenos Aires province, Argentina. **Information Processing in Agriculture**, v. 2, n. 2, p. 73–84, 2015.

MACDONALD, R. B.; HALL, F. G. Global crop forecasting. **Science**, v. 208, n. 4445, p. 670–679, 1980.

MONTEITH, J. L. Climate and the Efficiency of Crop Production in Britain [and Discussion]. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 281, n. 980, p. 277–294, 1977.

MOULIN, S.; BONDEAU, A.; DELECOLLE, R. Combining agricultural crop models and satellite observations: From field to regional scales. **International Journal of Remote Sensing**, v. 19, n. 6, p. 1021–1036, 1998.

NGUY-ROBERTSON, A. et al. Green leaf area index estimation in maize and soybean: Combining vegetation indices to achieve maximal sensitivity. **Agronomy Journal**, v. 104, n. 5, p. 1336–1347, 2012.

ORT, D. R.; LONG, S. P. Limits on yields in the Corn Belt. **Science**, v. 344, n. 6183, p. 484–485, 2014.

PENG, B. et al. Benefits of Seasonal Climate Prediction and Satellite Data for Forecasting U.S. Maize Yield. **Geophysical Research Letters**, v. 45, n. 18, p. 9662–9671, 2018.

PERALTA, N. et al. Mid-season high-resolution satellite imagery for forecasting site-specific corn yield. **Remote Sensing**, v. 8, n. 10, p. 1–16, 2016.

SAKAMOTO, T.; GITELSON, A. A.; ARKEBAUER, T. J. Near real-time prediction of U.S. corn yields based on time-series MODIS data. **Remote Sensing of Environment**, v. 147, p. 219–231, 2014.

SCHWALBERT, R. A. et al. Forecasting maize yield at field scale based on high-resolution satellite imagery. **Biosystems Engineering**, v. 171, p. 179–192, 2018.

SHANAHAN, J. F. et al. Use of Remote-Sensing Imagery to Estimate Corn Grain Yield. **Agronomy Journal**, v. 93, p. 583–589, 2001.

SHAO, Y. et al. An analysis of cropland mask choice and ancillary data for annual corn yield forecasting using MODIS data. **International Journal of Applied Earth Observation and Geoinformation**, v. 38, p. 78–87, 2015.

SHELESTOV, A. et al. Exploring Google Earth Engine Platform for Big Data Processing: Classification of Multi-Temporal Satellite Imagery for Crop Mapping. **Frontiers in Earth Science**, v. 5, p. 17, 24 fev. 2017.

SIBLEY, A. M. et al. Testing remote sensing approaches for assessing yield variability among maize fields. **Agronomy Journal**, v. 106, n. 1, p. 24–32, 2014.

STEINMETZ, S. et al. Spectral estimates of the absorbed photosynthetically active radiation and light-use efficiency of a winter wheat crop subjected to nitrogen and water deficiencie. **International Journal of Remote Sensing**, v. 11, n. 10, p. 1797–1808, 1990.

TUCKER, C. J.; HOLBEN, B. N.; ELGIN, J. H. Relationship of spectral data to grain yield variation. **Photogrammetric Engineering and Remote Sensing**, v. 45, n. 5\, p. 657–666, 1980.

WIEGAND, C.; RICHARDSON, A. Use of spectral vegetation indices to infer leaf area, evapotranspiration and yield: I. rationale. **Agronomy Journal**, v. 82, p. 623–629, 1990.

YOU, J. et al. **Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data**. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17). **Anais**...2017

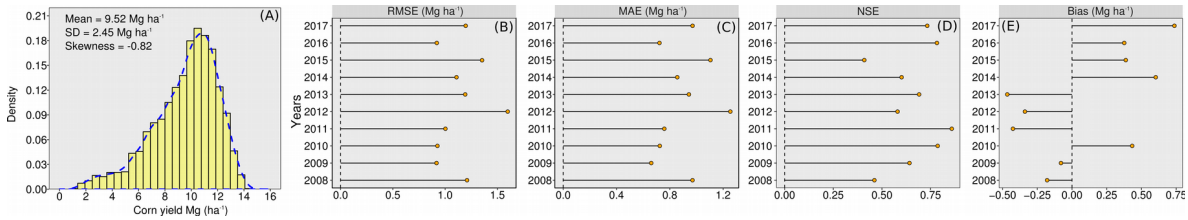
APÊNDICE A – TABELA DE EQUAÇÕES DOS ÍNDICES DE VEGETAÇÃO

Supplementary Table - Equations for the vegetation indices used in this study

Índices de Vegetação	Sigla	Equação
Normalized Difference Vegetation Index	NDVI	(Red - NIR) / (Red + NIR)
Green Normalized Difference Vegetation Index	GNDVI	(Green - NIR) / (Green + NIR)
Enhanced Vegetation Index	EVI	(NIR - Red) / (NIR + C1 x Red - C2 x Blue +L)
Normalized Difference Red Edge Index	NDRE, NDVIre	(Red-edge - NIR) / (Red - edge + NIR)

*Red representa a reflectância na região do vermelho, NIR representa a reflectância na região do infravermelho próximo, Blue representa a reflectância na região do azul, Green representa a reflectância na região do verde, Red-Edge representa a reflectância na região da borda do vermelho, L representa um o ajuste de fundo do dossel que trata da transferência de radiação não-linear para o NIR e vermelho através do dossel, C1, C2 são os coeficientes de resistência ao aerossol, esses coeficientes usam o comprimento de onda do azul para corrigir influências de aerossol na faixa vermelha, G representa o fator de ganho. Os coeficientes adotados no algoritmo MODIS-EVI são; L = 1, C1 = 6, C2 = 7,5 e G = 2,5.

APÊNDICE B – FIGURA SUPLEMENTAR 1 DO ARTIGO 1



Supplementary figure 1. (A) Observed yield data distribution. (B) Root-mean absolute error (RMAE). (C) Mean absolute error (MAE). (D) Nash–Sutcliffe model efficiency coefficient (NSE). (E) Bias coefficient for all the years considered in this study.

APÊNDICE C – TABELA SUPLEMENTAR 1 DO ARTIGO 3

Supplementary Table 1. Multiple linear regression models using Sentinel 2 full resolution for Rio Grande do Sul (RS) and Mato Grosso (MT).

Satellite description			RS model		MT model	
Band	Band name	Wavelength	Coefficient	P-value	Coefficient	P-value
Intercept	-	-	10.929022	< 2e-16	3.5819608	1.98e-06
Band 2	Blue	490 nm	-0.0001439	0.89075	0.0002143	0.868
Band 3	Green	560 nm	0.0025959	0.00113	0.0067544	< 2e-16
Band 4	Red	665 nm	-0.0027115	2.5e-12	0.0084745	< 2e-16
Band 5	Red Edge 1	705 nm	-0.0013806	3.2e-05	-0.0019428	5.10e-06
Band 6	Red Edge 2	740 nm	-0.0098994	< 2e-16	-0.0152953	< 2e-16
Band 7	Red Edge 3	783 nm	-0.0001205	0.61825	-0.0004666	0.0564
Band 8	NIR	842 nm	0.0088876	< 2e-16	0.0104068	< 2e-16
Band 8a	Red Edge 4	865 nm	-0.0010748	3.8e-08	-0.0001988	0.3580
Band 11	SWIR 1	1610 nm	0.0006389	0.053	0.0004448	0.2337
Band 12	SWIR 2	2190 nm	0.0003569	0.415	-0.0001865	0.7715

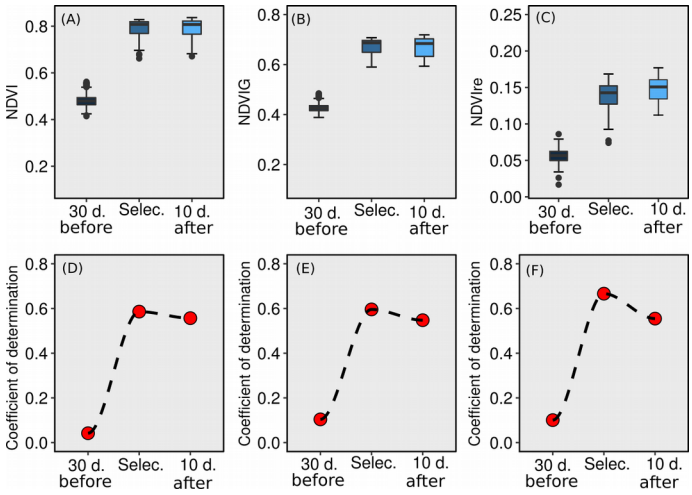
APÊNDICE D – TABELA SUPLEMENTAR 2 DO ARTIGO 3

Supplementary Table 2. Moran's I test to vegetation indexes (VI's) obtained from mid-season satellite imagery and yield monitor data.

Field	State	Season	Maize yield	NDVI	NDVIG	NDVIre
F1	RS	2016-2017	0.49***	0.48***	0.52***	0.48***
F2	RS	2016-2017	0.32***	0.35***	0.42***	0.30***
F3	RS	2016-2017	0.20***	0.22***	0.25***	0.26***
F4	RS	2016-2017	0.17***	0.15***	0.21***	0.12***
F5	RS	2016-2017	0.18***	0.19***	0.23***	0.20***
F6	RS	2016-2017	0.13***	0.16***	0.19***	0.21***
F7	MT	2016	0.25***	0.23***	0.31***	0.22***
F8	MT	2016	-0.05	0.03	0.07	0.04
F9	MT	2016	0.17***	0.21***	0.23***	0.28***
F10	MT	2016	0.24***	0.27***	0.31***	0.28***
F11	MT	2016	0.15***	0.19***	0.20***	0.18***
F12	MT	2017	0.21***	0.24***	0.26***	0.24***
F13	MT	2017	0.28***	0.26***	0.30***	0.24***
K1	KS	2016	0.08***	0.20***	0.18***	0.14***
K2	KS	2016	0.12***	0.15***	0.14***	0.09***
K3	KS	2016	0.14***	0.15***	0.16***	0.15***
K4	KS	2016	0.04***	0.25***	0.26***	0.18***
K5	KS	2016	0.06***	0.15***	0.18***	0.13***
K6	KS	2016	0.05***	0.17***	0.16***	0.14***

Notes: The statistically significant coefficients are indicated by asterisks, where * Significant at the alpha = 0.05 error level; ** Significant at the alpha = 0.01 error level; *** Significant at the alpha = 0.001 error level.

APÊNDICE E – FIGURA SUPLEMENTAR 1 DO ARTIGO 3



Supplementary figure 1. Boxplot showing NDVI (A), NDVIG (B) and NDVIre (C) range from different image date acquisition. The lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles). The upper whisker extends from the hinge to the largest value no further than 1.5 x IQR (inter-quartile range) from the hinge. Coefficient of determination (R^2) versus image date acquisition from a yield-NDVI (D), yield-NDVIG (E) and yield-NDVIre (F) relationship. Selec. = Selected image used to build the forecasting yield models. 30 d. before = Image acquired 30 days before the selected image. 10 d. after = Image acquired 10 days after the selected image.