**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

**A relação entre o tamanho das propriedades agrícolas e a produtividade no
Brasil: uma análise não paramétrica**

Alexandre Amorim de Souza Ferreira

Dissertação apresentada para obtenção de título de Mestre
em Ciências. Área de concentração: Economia Aplicada.

**Piracicaba
2018**

Alexandre Amorim de Souza Ferreira
Bacharel em Ciências Econômicas

**A relação entre o tamanho das propriedades agrícolas e a produtividade no Brasil:
uma análise não paramétrica**

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:

Prof. Dr. **ALEXANDRE NUNES DE ALMEIDA**

Dissertação apresentada para obtenção de título de Mestre
em Ciências. Área de concentração: Economia Aplicada.

Piracicaba
2018

Dados Internacionais de Catalogação na Publicação

DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Ferreira, Alexandre Amorim de Souza

A relação entre o tamanho das propriedades agrícolas e a produtividade no Brasil: uma análise não paramétrica / Alexandre Amorim de Souza Ferreira. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2018.

92 p.

Dissertação (Mestrado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. Relação entre área e produtividade 2. Propriedades agrícolas 3. Elasticidade produção 4. Redistribuição de terras 5. Regressão não paramétrica I. Título

DEDICATÓRIA

Dedico este trabalho aos meus pais Carlos Alberto de Souza Ferreira e Regina Martins Amorim de Souza Ferreira, ao meu irmão Guilherme Amorim de Souza Ferreira e a toda minha família, que sempre estiveram ao meu lado e aos meus amigos de Itapira e da República K-labouço, os quais sempre estiveram presentes em todos os momentos.

AGRADECIMENTOS

Em primeiro lugar agradeço aos meus pais, Carlos Alberto e Regina, pelo apoio, motivação, ajuda e por sempre acreditarem em mim. Agradeço ao meu irmão Guilherme, à República K-labouço e a todos os meus amigos, por sempre estarem presentes nos momentos de alegria e dificuldade além das energias, sempre positivas, enviadas a mim.

Agradeço a toda a minha família por ser tão unida e o pilar de todas as minhas conquistas. Ao meu orientador, o professor Alexandre Nunes de Almeida, pela paciência e orientação durante esta caminhada.

Aos professores e colaboradores do departamento, pelos ensinamentos e pelas oportunidades durante a graduação e a pós-graduação, assim como todos que de alguma forma fazem parte da ESALQ-USP, instituição que sempre terei na memória.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) que me auxiliaram financeiramente, por meio da bolsa de estudos do mestrado.

Por fim, e não menos grato, à Deus. Obrigado por tudo!

SUMÁRIO

RESUMO	6
ABSTRACT	7
LISTA DE FIGURAS	8
LISTA DE TABELAS	9
1 INTRODUÇÃO.....	11
2 REVISÃO DE LITERATURA	15
3 MOTIVAÇÃO PARA A REGRESSÃO NÃO PARAMÉTRICA	25
3.1 Uso da regressão não paramétrica para a análise da relação entre o tamanho das propriedades agrícolas e a produtividade no mundo	27
3.2 A realidade brasileira.....	30
4 MATERIAIS E MÉTODOS.....	33
4.1 A função de produção Cobb-Douglas e a relação entre a produtividade e o tamanho das propriedades agrícolas.....	33
4.2 O modelo de regressão <i>kernel</i> multivariado não paramétrico	36
4.3 O estimador constante local (LCLS- <i>local constant linear squares</i>)	40
4.4 Estimador linear local (LLLS- <i>Local linear least squares</i>).....	42
4.5 Estimador quadrático local (LQLS- <i>Local quadratic least squares</i>)	44
4.6 Base de dados e estatística descritiva	45
4.7 Variáveis utilizadas no modelo econométrico.....	46
4.8 Estatística descritiva das variáveis utilizadas no modelo econométrico	48
5 RESULTADOS	57
6 ROBUSTEZ DOS RESULTADOS.....	65
7 CONCLUSÕES	87
REFERÊNCIAS	89

RESUMO

A relação entre o tamanho das propriedades agrícolas e a produtividade no Brasil: uma análise não paramétrica

A análise de regressão *kernel* não paramétrica desconsidera qualquer influência das formas funcionais geralmente empregadas em análises de regressões paramétricas, permitindo os dados “falarem por si mesmos”. Enquanto os estimadores paramétricos são considerados globais, os *kernels* não paramétricos usam uma amostra de dados próximas (definida pela largura da janela) a um ponto para ajustar a estimação, o que permite focar em peculiaridades locais dos dados. Ambas as análises foram aplicadas aos dados do Censo Agropecuário de 2006 realizado pelo IBGE, agregados municipalmente e em dezessete faixas de áreas, para estimar uma função de produção com o objetivo de estabelecer a relação entre o tamanho das propriedades agrícolas e o valor da produção por hectare (produtividade). A relação constatada foi inversa, porém a análise local feita pelos estimadores *kernels* explicitou uma relação direta entre as elasticidades de produção dos insumos e o tamanho das propriedades agrícolas, o que não justifica uma política de redistribuição de terras no sentido do aumento da produtividade. Além disto, análises gráficas contra fatuais (que manteve os insumos, exceto a área, constantes em seus valores médios) mostraram que a relação não é linear, não é monotônica, e difere dentre as regiões, o que é um desafio para a elaboração de políticas de redistribuição de terras.

Palavras-chave: Relação entre área e produtividade; Propriedades agrícolas; Elasticidade produção; Redistribuição de terras; Regressão não paramétrica.

ABSTRACT

The relationship between farm size and productivity in Brazil: a nonparametric analysis

Nonparametric kernel regression analysis disregards any influence of the functional forms commonly employed in parametric regression analyzes, allowing the data to "speak for itself." While parametric estimators are considered global, nonparametric kernels use a sample of nearby data (defined by the bandwidth) at a point to adjust the estimation, which allows focusing on local peculiarities of the data. Both analyzes were applied to data from the 2006 IBGE Census of Agriculture, aggregated in municipalities and in seventeen areas, to estimate a production function with the objective of establishing the relationship between the size of agricultural properties and the value of production by hectare (productivity). The observed relationship was reversed, but the local analysis made by the kernels estimators explained a direct relationship between the elasticities of production of the inputs and the size of the agricultural properties, which does not justify a policy of redistribution of land in order to increase productivity. In addition, graphical analyzes against factors (which kept the inputs, except the area, constant in their mean values) showed that the relationship is not linear, is not monotonic, and differs among regions, which is a challenge for the elaboration of land redistribution policies.

Keywords: Relationship between farm size and productivity; Agricultural properties; Elasticities of production; Redistribution of land; Nonparametric regression.

LISTA DE FIGURAS

Figura 1. Representação da maldição da dimensionalidade em uma e duas dimensões.....	26
Figura 2. Relação entre o logaritmo do valor da produção (em mil reais) por hectare pela área (em hectares) para o Brasil para os estimadores LCLS e LLLS	62
Figura 3. Relação entre o logaritmo natural do valor da produção (em mil reais) por hectare e a área (em hectares) para o Brasil e o Sul mantendo-se fixas as outras variáveis (LCLS).....	81
Figura 4. Relação entre o valor da produção (em mil reais) por hectare pela área (hectares) para o Centro Oeste e o Nordeste mantendo-se as outras variáveis fixas (LCLS)	82
Figura 5. Relação entre o valor da produção (em mil reais) por hectare pela área (em hectares) para o Norte e o Sudeste mantendo-se as outras variáveis fixas (LCLS).....	83
Figura 6. Relação entre o valor da produção (em mil reais) por hectare pela área (em hectares) para o Brasil e regiões mantendo-se as outras variáveis fixas (LLLS)	84

LISTA DE TABELAS

Tabela 1. Estatística descritiva por regiões considerando municípios como observações agregadas.	49
Tabela 2. Estatística descritiva por região considerando as fazendas representativas como observações agregadas.....	51
Tabela 3. Estatística descritiva por faixa de área.....	52
Tabela 4. Regressão paramétrica da produtividade da terra admitindo a função de produção Cobb-Douglas com as regiões representadas por variáveis <i>dummies</i>	57
Tabela 5. Regressão paramétrica da produtividade da terra admitindo a função de produção Cobb-Douglas e regiões separadas	59
Tabela 6. Regressão <i>kernel</i> não paramétrica utilizando os três estimadores com o valor da produção por hectare como variável dependente	61
Tabela 7. Regressão paramétrica utilizando a função de produção Cobb-Douglas e as fazendas representativas como observações.....	66
Tabela 8. Regressão paramétrica utilizando a função de produção Cobb-Douglas para o Brasil e regiões, considerando as fazendas representativas como observações.....	67
Tabela 9. Regressão <i>kernel</i> não paramétrica utilizando as fazendas representativas como observações e 3 estimadores para o Brasil	69
Tabela 10. Regressão <i>kernel</i> não paramétrica considerando as fazendas representativas como observações para a região Sudeste.....	71
Tabela 11. Regressão <i>kernel</i> não paramétrica considerando as fazendas representativas como observações para a região Sul.....	73
Tabela 12. Regressão <i>kernel</i> não paramétrica considerando as fazendas representativas como observações para a região Centro Oeste	75
Tabela 13. Regressão <i>kernel</i> não paramétrica considerando as fazendas representativas como observações para a região Norte.....	77
Tabela 14. Regressão <i>kernel</i> não paramétrica considerando as fazendas representativas como observações para a região Nordeste.....	79

1 INTRODUÇÃO

A relação entre o tamanho das propriedades e a produtividade da terra agrícola é um tema antigo e amplamente discutido na literatura econômica. Benjamin (1995) mostra que se ela for direta, um aumento na área produtiva irá promover um aumento na produtividade de maior magnitude que o aumento na área; quando inversa, um aumento na área produtiva terá, como consequência, um aumento de produtividade menor que o da área (ou ainda uma redução de produtividade). Em outras palavras, quando a relação é inversa, as menores fazendas têm maior produção por medida de área e vice e versa.

O conhecimento empírico, em países com menor nível de desenvolvimento, sugere que esta relação seja negativa (FEDER, 1985; CHEN et al., 2011; ALI e DEININGER, 2015). Benjamin (1995), assim como Chen et al. (2011), afirmam que Chayanov (1926) observou pela primeira vez este fenômeno para a Rússia. No entanto a primeira referência sobre o assunto é creditada a Sen (1962).

Após a primeira referência moderna, houve um crescente interesse sobre o assunto. Na década de 70, autores como Sen (1975), Berry e Cline (1979) escreveram livros¹ que se tornaram referências importantes. A partir da década de 80 observou-se um grande aumento de publicações² e de estudos empíricos sobre o tema.

Várias explicações foram tecidas e testadas ao longo destas décadas, mas nenhuma delas ainda foi suficiente para explicar completamente a hipótese da relação inversa (VERSCHELDE et al., 2013).

Primeiro, destaca-se as imperfeições nos mercados de trabalho (FEDER, 1985; HELTBERG, 1998; ALI; DEININGER, 2015), de crédito (BINSWANGER; ROSENZWENG, 1986; HELTBERG, 1998; ASSUNÇÃO; GATHAK, 2003), de seguros (BINSWANGER et al., 1995) e de terras (HELTBERG, 1998) que são apontadas como importantes razões para a relação inversa. Estas falhas, ou ainda a ausência de algum destes mercados, geram distorções na alocação de recursos nas fazendas, fazendo com que ela não seja ótima e implique em ineficiências (VERSCHELDE et al., 2013).

Os trabalhadores contratados não partilham os mesmos riscos, são menos motivados e se esforçam menos que os familiares, o que gera um alto custo de supervisão do trabalho

¹ Sen (1975) escreveu o livro “Employment, Technology and Development” enquanto que Berry e Cline (1979) o “Agrarian structure and productivity and developing countries”.

² Os trabalhos sobre o assunto serão melhores abordados na seção de revisão de literatura.

(FEDER, 1985; HELTBERG, 1998). As instituições financeiras, devido a assimetria de informações e ao risco da agricultura, racionam o crédito (CARTER; WIEBE, 1990) e requerem garantia para a realização de empréstimos, dificultando o acesso aos produtores com menores áreas (BINSWANGER; ROSENZWENG, 1995). O mercado de seguros geralmente atua contra eventos climáticos. Binswanger e Rosenzweng (1995) destacam a falta deste mercado para risco de preço para países em desenvolvimento e a priorização deste mercado para as maiores propriedades. Heltberg (1998) explicita problemas no mercado de aluguel de terras (se o contrato é compartilhado, o arrendatário tende a usar menos insumos e subestimar a renda; se o contrato é pago pós colheita, há o risco de quebra de safra e o não pagamento do contrato; se o aluguel é pago antes da colheita, pode ocorrer falta de liquidez e impedir o acesso a crédito e reduzir o potencial de investimento na produção) e de venda (difícil acesso a financiamentos para a compra de terras restringe o acesso às terras).

Outra explicação bastante comum na literatura é a relação inversa causada por vieses de variáveis omitidas, como por exemplo, a qualidade do solo (BENJAMIN, 1995; CARLETTO et al., 2013). Diferenças na qualidade do solo geram diferenças na produção. O controle para esta variável não foi capaz de explicar a relação inversa para os trabalhos de Barret et al. (2010), Carletto et al. (2013), mas inverteu a relação para a China (CHEN et al., 2011).

As características das famílias, proprietários e trabalhadores também são consideradas variáveis omitidas (ASSUNÇÃO; GATHAK, 2003; ASSUNÇÃO; BRAIDO, 2007). Assunção e Braido (2007) investigaram a influência destas variáveis para Índia, cujo resultado demonstrou que elas não explicaram a relação em questão. Além disto, evidências também sugerem que as diferenças entre as habilidades gerenciais não totalmente explicam a relação inversa (VERSCHELDE et al., 2013).

Questões metodológicas são a terceira explicação tradicional encontrada na literatura. Lipton (2010) afirma que as propriedades de diversos tamanhos geralmente são consideradas réplicas lineares uma das outras, o que pode gerar viés nas estimações, visto que elas não são. As maiores propriedades utilizam diferentes tecnologias que as menores e o uso de insumos varia com a escala da produção e o tamanho das propriedades. Carletto et al. (2013) investigaram se os erros de mensuração e arredondamentos das áreas poderiam ser a causa da relação inversa para Uganda, cuja conclusão foi negativa.

Entender a relação entre o tamanho das propriedades agrícolas e a produtividade, assim como suas possíveis explicações encontradas na literatura, é de suma importância

porque colabora para que políticas de redistribuição de terras sejam planejadas. Se a relação inversa é observada, então uma redistribuição de terras é importante para aumentar a produtividade da agricultura, o que pode justificar, por exemplo, uma política de reforma agrária (DELOLAIKAR, 1981; CORNIA, 1985; CHEN et al., 2011). Além disso, Helfand et al. (2015) enaltecem que o Brasil apresentava políticas relacionadas às grandes propriedades focadas fortemente ao mercado, que são fomentadas pelo Ministério da Agricultura, e também às pequenas propriedades (propriedades familiares) pelo Ministério do Desenvolvimento Agrário. Esta dualidade de políticas faz do Brasil um interessante objeto para este estudo, e somada às importâncias de se entender a relação entre o tamanho das propriedades agrícolas e a produtividade das mesmas, justificam esta dissertação.

Portanto, este trabalho tem como objetivo analisar a relação entre o tamanho das propriedades agrícolas brasileiras e a produtividade das mesmas através da estimação de uma função de produção Cobb-Douglas paramétrica e não parametricamente. A função Cobb-Douglas é preferida em relação às demais porque ela é mais parcimoniosa, o que é benéfico para o processo de identificação da estrutura dos dados feito pela econometria não paramétrica. Para isso, utilizou-se os dados coletados pelo Instituto Brasileiro de Geografia e Estatística (IBGE) através do Censo Agropecuário do ano de 2006.

A utilização do método não paramétrico³ será uma grande contribuição, visto que não foram encontrados trabalhos que utilizem deste método para este tipo de análise na literatura brasileira. Ele não utiliza nenhuma assunção paramétrica para a identificação da estrutura dos dados, ou seja, permite, segundo Eubank (1999 apud ALMEIDA, 2016) “os dados falarem por si mesmos”. Além disto, Verschelde et al. (2013) citam autores que sugerem o uso deste tipo de regressão para analisar esta relação.

Este trabalho será apresentado da seguinte forma: A seção seguinte trará a revisão de literatura sobre o tema; a terceira a motivação para o uso da regressão não paramétrica e a aplicação dela para a análise da relação entre a área e a produtividade das propriedades agrícolas; na quarta seção será abordada a metodologia; a quinta serão apresentados os dados e as estatísticas descritivas deles; na sexta os resultados e os devidos comentários serão tecidos; na sexta um exercício de robustez será feito; e por fim, na sétima e última seção, as conclusões serão apresentadas.

³ Este método será melhor explorado em seções posteriores.

2 REVISÃO DE LITERATURA

Este tema já foi amplamente discutido e testado na literatura econômica empírica e teórica através de diferentes modelos, bancos de dados e variáveis que resultaram em uma gama de hipóteses e conclusões. Por isso, sente-se necessário trazer uma revisão detalhada destes trabalhos com o intuito de ilustrar como os autores analisaram de diferentes maneiras esta mesma relação, assim como suas conclusões para diversos países, que é apresentada nesta seção. Na próxima, apresenta-se a motivação para o uso da regressão não paramétrica, seguida dos trabalhos que fazem uso deste método para este propósito (análise da relação entre a área e a produtividade das propriedades agrícolas), e dos estudos brasileiros relacionados com tema.

A versão moderna da discussão sobre a controvérsia entre o tamanho das propriedades e a produtividade agrícola surgiu após a publicação dos resultados do “*Indian Farm Management Studies*” na década de 50 (CORNIA, 1985). À Sen (1962) é creditada a referência moderna pioneira em constatar a relação inversa entre o tamanho das propriedades e a produtividade agrícola. Ele fez uma interpretação da relação que foi observada na agricultura indiana neste estudo realizado pelo Ministério da Alimentação e da Agricultura, nos anos de 1954 a 1956, para seis regiões do país. Em todas as áreas estudadas, um grande número de propriedades tem perdas significantes ao se considerar um valor salarial ao trabalho familiar, o que o autor relaciona com a pobreza deste setor. A lucratividade (tida como o resultado das receitas subtraídos os custos) aumenta conforme o tamanho das áreas produtivas aumentam, fato que ele explica pela alta rentabilidade do algodão em Punjab, geralmente plantado nestas grandes propriedades; pela indivisibilidade de bens de capital e as consequências da economia de larga escala em Uttar Pradesh; pelo aumento da quantidade de uso de insumos menor que o aumento na produção em Madras. E por fim, ele constata uma relação inversa entre a produtividade e a área produtiva, e atribui como causa o maior uso de insumos e trabalho familiar por área nas menores propriedades.

A utilização de funções de produção para a análise empírica das relações entre insumos e produtos é comum na literatura e passou a ser usual para investigar a relação entre a produtividade e o tamanho das propriedades agrícolas. Ulveling e Fletcher (1970) citam Walters (1963) o qual explicita que muitas tentativas foram feitas para encontrar formas algébricas de funções que sejam válidas teoricamente, e empiricamente usuais para descrever as relações entre insumo e produção de forma adequada. Eles afirmam que não há uma forma

funcional perfeita, e que cada forma funcional alternativa tem suas vantagens e desvantagens que devem ser consideradas ao se escolher a função a ser utilizada em um modelo.

Dolalakar (1981) utilizou dados do Censo Agrícola de 1970 e 1971, de um estudo realizado pela Universidade de Nehru e da Estatística de Fertilizantes do ano de 1972 para testar as hipóteses de que a relação inversa é válida para toda a agricultura da Índia e se ela também é válida com a presença do progresso tecnológico. Ele fez uso de uma função de produção para cada distrito, estimou os coeficientes através do método dos Mínimos Quadrados Ordinários e concluiu que ambas as hipóteses são confirmadas pelos dados, os quais não foram capazes de explicar o porquê das diferenças relativas de renda para as pequenas fazendas com a presença do progresso tecnológico.

Cornia (1985) utilizou dados da Organização das Nações Unidas para Alimentação e Agricultura (FAO) para analisar a relação entre os insumos, a produtividade do trabalho, da terra, a renda e o retorno de escala para 15 países em desenvolvimento entre os anos de 1973 e 1979. Através de análises de regressões ele estimou funções de produção pelo método dos mínimos quadrados ordinários para cada país. Foi concluído que o rendimento da terra é maior nas menores propriedades; que há retornos decrescentes de escala somente em países com abundância de terras na África, Oriente Médio e México e, nestes casos, a relação inversa pode ser atribuída aos retornos decrescentes de escala somadas ao alto uso de insumos pelos pequenos; e por fim que a elasticidade de produção da terra varia inversamente a área per capita.

Através de um modelo teórico de equilíbrio geral para maximização do lucro, diferente das funções de produção observadas até agora, Feder (1985) buscou explicações para a relação inversa. As explicações dependem de três proposições incorporadas no modelo: O trabalho contratado é mais eficiente quando sujeita a maior supervisão; os familiares são mais motivados e, portanto, exercem o papel de supervisionar os contratados; e o capital de giro é positivamente relacionado com o tamanho das propriedades. O modelo mostrou que há uma relação positiva entre a produtividade dos assalariados e a intensidade da supervisão. Quando imperfeições no mercado de crédito são consideradas, os resultados são diferentes devido a diferenças no capital de giro, gerada por esta falha, dentre os produtores. E por fim, o subsídio a insumos (não relacionados a mão de obra) tendem a reduzir o trabalho total no curto prazo, com a presença da supervisão dos mesmos.

Bhalla (1988) desenvolveu uma metodologia para mensurar a qualidade da terra para as propriedades agrícolas da Índia. Através dos dados disponibilizados pelo Conselho

Nacional de Pesquisa em Economia Aplicada (NCAER) para os anos de 1976 e 1977, estimou uma função de produção que fez uso de informações relativas ao solo, irrigação, número de parcelas e região das propriedades. Os dados usados possibilitaram criar um índice capaz de comparar a qualidade da terra em diferentes fazendas. As principais conclusões do modelo foram que a irrigação não é uma *proxy* adequada para a qualidade da terra; e que a correlação entre o índice de qualidade do solo desenvolvido e o tamanho das propriedades é negativa, resultando em importantes implicações na avaliação da distribuição de terras e da pobreza rural.

Semelhante a Feder (1985), Carter e Wiebe (1990) utilizaram um modelo de equilíbrio geral computável para analisar como os padrões de acesso ao capital influenciam os padrões de alocação de recurso e produtividade para o setor dos pequenos proprietários no Quênia. Eles observaram que o acesso a capital tem um importante papel na produtividade agrícola e no mercado de terras, privilegiando os maiores produtores.

Através de um modelo teórico de dois períodos para maximizar a utilidades das famílias, Barret (1990) demonstrou que a renda face a incerteza de preços, somada a má distribuição de terras nos países de baixa renda, são suficientes para explicar a relação inversa, desde que os pequenos produtores sejam avessos ao risco. Ele também utilizou dados de Madagascar para estilizar sua hipótese, no entanto, a análise empírica não é tão relevante porque os dados não foram coletados para este propósito, o que não permite uma robustez dos coeficientes. Ele considera em seu modelo uma homogeneidade nos solos e um mercado de trabalho competitivo, demonstrando que, mesmo sob estas condições, a relação inversa persiste.

Ao desconsiderar as variáveis omitidas, Benjamin (1995) constatou, através da análise de regressão de uma função de produção, que a relação inversa e a sensibilidade salarial são divergentes da interpretação da produção dos agricultores e da demanda de trabalho geradas pela maximização dos lucros. Devido à ausência de dados destas variáveis omitidas, ele emprega a técnica das variáveis instrumentais que, além de desaparecer com a relação inversa, produziu elasticidades salariais mais convergentes com a teoria econômica e explicitou a importância do viés da variável omitida. Para testar se a qualidade da terra pode ser interpretada como a variável omitida, o autor fez uso de um modelo que comparou os vieses (através da estimativas de distância mínima⁴) da estimação, pelo método dos Mínimos

⁴ Segundo Beran (1984) a estimativa de mínima distância são estimativas dos parâmetros que buscam minimizar a distância entre os dados e o modelo estimado.

Quadrados Ordinários, de três configurações diferentes para demonstrar como a qualidade da terra pode afetar a produção: Primeiro ele considera que a qualidade da terra tem um efeito neutro; por segundo que a produtividade da terra é aumentada pela qualidade da terra; e por fim, que o trabalho é aumentado pela qualidade da terra. Apesar da rejeição formal, o modelo leva a conclusão de que a qualidade da terra é uma fonte plausível para a relação inversa.

A relação inversa tem sido frequentemente observada em trabalhos anteriores. Em grande parte deles, ela é atribuída a maior produtividade do trabalho familiar, o qual as pequenas propriedades fazem maior uso e imperfeições de mercado. Frente a isto, Binswanger et al. (1995) procuraram melhor entender as inequidades dos mercados de venda e aluguel de terras em países em desenvolvimento. Para isto, eles dividiram o estudo em três partes: A primeira faz uma análise histórica que aborda a emergência dos direitos de propriedade, o aluguel e tributos de camponeses, sucessos e falhas da reforma agrária; a segunda traz controvérsias analíticas que aborda a relação inversa entre o tamanho das propriedades agrícolas e a produtividade, efeitos do crédito, distorções políticas e incentivos no mercado de terras; e por fim discute tópicos relacionados a política de terras (taxas, registros, fragmentação e consolidação). A segunda é a com maior relação ao presente trabalho. Nela eles trazem informações empíricas sobre a relação inversa em alguns países em desenvolvimento (inclusive o Brasil). Os autores destacaram por que o ganho de escala no processo agrícola de produções de uma única cultura totalmente voltada ao mercado (caso da cana de açúcar no Brasil), as vantagens das grandes fazendas de utilizar insumos tecnológicos (como máquinas) e as diferenças de acesso a crédito a difusão do risco não são suficientes para explicar a relação inversa. A superioridade de produção das pequenas é destacada neste trabalho, mas a importância de mais trabalhos sobre o assunto é requisitada.

A evidência da relação inversa é constatada para o Paquistão por Heltberg (1998). Foi usado dados da Pesquisa Rural do Paquistão para montar um painel entre os anos 1986 e 1987 a 1990 e 1991 e estimar uma função de produção com efeitos fixos e aleatórios. Cinco diferentes variáveis dependentes foram utilizadas: Valor adicionado por área utilizada; retorno por área total; lucro das culturas por área utilizada; crédito por área utilizada (todos em rupias); e trabalho (homens por dia) por área utilizada. As variáveis independentes foram: A área operada; a área total; o trabalho familiar; o risco; e o vetor de outras características. Uma relação em forma de U (a variável área foi considerada polinomial de grau 3) foi obtida entre o valor adicionado por área e a área operada, que é a de maior interesse para esta dissertação devido a simetria entre as variáveis dependentes. A relação foi inversa para as propriedades

com até 16 acres, que representam 90% da amostra. Este formato irregular demonstra que não há um padrão definido para a relação, ou seja, um método não paramétrico seria adequado para essa análise. O crédito apresentou formato semelhante ao valor adicionado; o trabalho familiar reduziu com a área operada; enquanto que o lucro apresentou uma relação positiva. Por fim, a relação inversa não foi explicada pelo viés das variáveis omitidas.

Dorward (1999) aplicou duas diferentes abordagens analíticas, já observadas em trabalhos anteriores, para analisar a relação entre o tamanho das propriedades e a produtividade para o Malawi. A primeira, a qual se dá maior enfoque devido a convergência metodológica com este trabalho, é através da análise de regressão para uma função de produção; a segunda, um modelo de programação linear para o comportamento das fazendas, para analisar o efeito das restrições de terra e capital na relação. Foi utilizado dados agregados entre os anos de 1980 e 1985, divididos em sete categorias de tamanho das propriedades e para a região específica de Mzimba, localizada ao norte do Malawi. A variável dependente da função de produção foi construída pela média da multiplicação da produção pelo preço das mesmas de todos os anos, subtraído os gastos com insumos por hectare, dividida pela média das áreas produtivas, e as independentes foram as médias das áreas cultivadas por parcela e variáveis *dummies* para cada categoria de tamanho de cada ano, através de quatro formas funcionais (quadrática, cúbica, log-linear e Cobb-Douglas). A que obteve melhor ajuste foi a Cobb-Douglas, com as variáveis explicativas sendo lineares e quadráticas. Observa-se duas semelhanças com este trabalho: a divisão dos dados agregados em categorias de tamanhos das propriedades e a utilização da função de produção no formato Cobb-Douglas. O autor constatou uma relação positiva para ambos os casos e atribuiu como explicação a pobreza, as restrições de capital e consumo, a utilização de sistema de cultivo baseado nas chuvas e a não existência de mercados de terras e crédito.

Assunção e Ghatak (2003), através de um modelo teórico, buscaram explicações alternativas para a relação inversa às falhas de mercado somadas aos retornos decrescentes de escala para a terra e insumos. Estas explicações se basearam na escolha de ocupação endógena e nas diferenças de habilidades agrícolas dos proprietários. A principal suposição do modelo é a não existência do mercado de crédito e que a média de tamanho da propriedade dos proprietários mais habilidosos são menores. A resposta foi que a relação inversa pode ser atribuída a imperfeições no mercado de crédito e nas diferenças de habilidades agrícolas dos proprietários. Além disto, eles afirmam que o custo de oportunidade de camponeses habilidosos se tornarem trabalhadores assalariados (fora das propriedades agrícolas) é alto.

Como consequência, a menos que a análise econométrica faça rigoroso controle destas habilidades, um viés irá extrapolar a verdadeira relação entre a produtividade e o tamanho das propriedades.

Lamb (2003) buscou entender o papel da qualidade da terra, do mercado de trabalho e erros de mensuração na relação inversa em questão para a Índia. Para tal, é estimada uma função de produção com a medida de produtividade como variável dependente e a área, insumos e outras características importantes nas independentes. Vale ressaltar que quando o coeficiente da área estimado for menor que 1 a relação inversa é considerada, mas quando a medida de produtividade for dividida pela área a relação inversa é considerada quando este coeficiente for menor que 0 (HELTBERG, 1998). Os dados foram retirados do Instituto de Pesquisa de Culturas Internacional para os Trópicos Semiáridos (ICRISAT) para diferentes regiões da Índia entre os anos de 1975 e 1985. Uma análise de dados em painel com efeitos fixos e aleatórios foi empregada. A variável dependente foi o lucro das propriedades e a demanda por trabalho. O autor insere nas regressões informações sobre a qualidade da terra para testar a influência dela na relação; características dos mercados de trabalho e de terra para testar a hipótese da influência das imperfeições do mercado de trabalho na relação inversa; e testa se erros de mensuração exercem influência através do uso de variáveis instrumentais para corrigi-los. A relação inversa não foi rejeitada e as principais inferências foram que a qualidade da terra e as falhas de mercado, especialmente no efeito aleatório, explicam bastante da relação inversa, e ressalta o importante papel dos erros de mensuração, especialmente nos efeitos fixos.

Estes mesmos dados foram utilizados por Assunção e Braidó (2007) para avaliar se as características específicas dos produtores e suas famílias podem ser a causa da relação inversa. A diferença é que eles exploraram somente os dados a níveis das parcelas utilizadas para a agricultura, enquanto Lamb (2003) utilizou dados agregados das fazendas. Em Lamb (2003), há uma observação de que é usual na literatura utilizar o valor da produção por unidade de área como medida de produtividade. Destaca-se este ponto porque a variável valor de produção por unidade de área é tomada como medida de produtividade neste trabalho. Foram estimadas diferentes funções de produção log-linear através do método dos Mínimos Quadrados Ordinários, cujo o valor da produção (em rupias) por área (acres) é a variável dependente. Primeiro a função foi estimada sem nenhum controle para as áreas utilizadas e em seguida foram adicionados controles relativos ao valor da terra, uso de irrigação e tipo de solo. Quando há o controle, a relação é suavizada, o coeficiente da área utilizada passa de -0,3

para -0,16, o que sugere que as menores parcelas têm mais atributos melhores para a produtividade. Em seguida, características relacionadas aos produtores (e suas famílias) e à área total produtiva foram adicionadas a regressão, que foi estimada através do método dos efeitos fixos. O número de membros adultos das famílias que trabalham nas propriedades foi usado para avaliar a hipótese do custo de supervisão do trabalho e o número de parcelas de cada propriedade para a hipótese do aumento do custo de supervisão devido a dispersão geográfica (considerou os custos similares entre os produtores que tem o mesmo número de parcelas). A relação inversa foi observada em todos os casos, além disso, as características relacionadas às famílias não explicam a relação, a qual, segundo o autor, está mais relacionada com as características das parcelas.

Kimhi (2006) fez uso de um modelo com dois estágios para analisar a relação entre a produtividade e o tamanho das parcelas da Zâmbia para os anos de 1993 e 1994. No primeiro estágio, ele estima a quantidade de terra alocada para a produção de milho. No segundo uma função de produção utilizando a quantidade estimada no primeiro, que devido a correlação do termo de erro da segunda regressão com a primeira, um modelo *tobit* é empregado. O autor também explicita que a forma mais utilizada na literatura de medida de produtividade para este tipo de análise (tamanho das propriedades e produtividade) é o valor da produção por unidade de área. A relação encontrada é positiva quando não é feita correção para endogeneidade da principal variável explicativa (tamanho das parcelas) e negativa quando esta correção é feita. Além disso, um padrão não linear entre a medida de produtividade e o tamanho das parcelas é observado. O emprego de uma regressão não paramétrica colaboraria para melhor entender o formato desta relação não linear, assim como para captar se a relação é uniforme ou não dentre os tamanhos das parcelas.

Barret et al. (2010) utilizou dados de Madagascar para o ano de 2002 para analisar duas explicações convencionais encontradas na literatura para a relação inversa: Imperfeições de mercado e omissão da qualidade do solo. Para isso, dois modelos foram empregados: A estimação de uma função de produção Cobb-Douglas com a medida de produtividade sendo a produção de milho (em quilogramas) por área cultivada (ares) e como variáveis independentes a área cultivada das parcelas (em ares) e o vetor com outras informações relacionadas com a produção (incluindo capital e trabalho); e a estimação de uma função hedônica de preços para o valor da terra. Atenta-se à primeira devido a maior simetria com este trabalho. A estratégia empírica utilizada é semelhante a Assunção e Braidó (2007), e consiste em testar a relação inversa sem nenhum controle, controlar para efeitos fixos das famílias, para a qualidade do

solo e ambos os controles simultaneamente. A relação inversa foi constatada e foi concluído que as imperfeições de mercado explicaram um terço da relação para Madagascar, enquanto que, com ou sem o controle de solo, o coeficiente de interesse continuou o mesmo, ou seja, não afetou a relação inversa.

A estimação de uma função de produção com e sem o uso de variáveis instrumentais é feita para analisar os impactos das características não observadas relativas à terra para a China por Chen et al. (2011). Foi construído um painel desbalanceado, estimado considerando efeitos fixos e aleatórios. Os autores consideraram a quantidade produzida de grãos (em toneladas) como variável dependente e como independentes a área (medida em logaritmos de Mu), o trabalho e características relacionadas às propriedades rurais. Em um primeiro momento, sem a aplicação das variáveis instrumentais, o coeficiente de interesse (da área) é de 0,841 para o efeito aleatório e 0,742 para o fixo. Como a quantidade produzida de grãos não foi mensurada em toneladas por Mu, e os coeficientes da variável área foram significativamente menores que 1, uma relação inversa pode ser observada (HELTBERG, 1998). Quando as variáveis instrumentais são aplicadas, estes coeficientes passam a ser bem próximos de 1 em ambos os casos. A fim de dar robustez aos resultados, um maior conjunto de instrumentos é utilizado para a mesma regressão, cujos resultados são bem próximos. A conclusão de Chen e colaboradores (2011) foi que, ao controlar para a qualidade do solo, a relação entre a produção de grãos e a área produtiva desaparece, logo a relação entre o tamanho das propriedades agrícolas e a produção de grãos para a China são proporcionais.

Com o objetivo de averiguar o impacto dos erros de mensuração sobre o tamanho das propriedades agrícolas, Carletto et al. (2013) utilizaram dados da Uganda para estimar dois modelos. A base de dados contou com a área das propriedades reportadas pelos proprietários e mensuradas por GPS (Sistema de Posicionamento Global), o que viabilizou o trabalho. O primeiro modelo teve como variável dependente o erro entre as medidas, e como variáveis independentes o tamanho das propriedades mensuradas pelo GPS e características dos proprietários e das parcelas, para analisar os possíveis determinantes dos erros. No segundo, duas funções de produção foram estimadas: Uma com a medida da área reportada pelos proprietários e a outra pelo GPS. As diferenças entre os resultados foram analisadas. A renda líquida agrícola por área em logaritmos (em dólares por área medida pelo GPS e pelos proprietários) foi a variável dependente, enquanto que as independentes foram o tamanho das propriedades medidas de ambas as formas e as características relativas às propriedades e produtores, todas divididas pelas áreas e em logaritmos. O coeficiente de interesse

(coeficiente da área) passou de -0,236, quando a área foi reportada pelos proprietários, a -0,374 pelo GPS, ou seja, quando a área é mensurada através do GPS, a relação passou a ser mais negativa, descartando a hipótese de a mesma ser um artefato estatístico gerado por erros de mensuração para a Uganda.

Esta seção ilustrou os estudos feitos sobre a relação entre o tamanho das propriedades agrícolas e a produtividade. Em grande parte da literatura analisada o uso da análise de regressão para estimar uma função de produção foi empregado para estabelecer essa relação. Além disso, a proposta de o valor da produção por unidade de área ser usado como medida de produtividade também foi comum dentre os autores. O uso da regressão não paramétrica por sua vez, será apresentado na próxima seção, após o uso do método não paramétrico ser justificado para este fim.

3 MOTIVAÇÃO PARA A REGRESSÃO NÃO PARAMÉTRICA

O uso de funções de produção para analisar a relação entre insumos e produtos tem sido feito há anos. A decisão sobre a sua forma funcional é de suma importância porque implica em suposições a priori à estimação dos dados no modelo (ULVELING; FLETCHER, 1970). Esta questão é crucial para este tema porque pode gerar vieses na estimação da relação entre a produtividade e a área (ULVELING; FLETCHER, 1970; DORWARD, 1999); preestabelece a direção dos retornos de escala (ULVELING; FLETCHER, 1970); não capta as não linearidades da relação (HELTBERG, 1998; DORWARD, 1999; KIMHI, 2006); e gera uma função de produção muito mais simples que a real (HENDERSON; PARMETER, 2014).

A principal vantagem do método *kernel* não paramétrico é permitir estimar uma função de produção sem predefinir a sua forma funcional (HENDERSON; PARMETER, 2014), ou seja, como na citação de Eubank (1999) feita por Almeida (2016) “permitir os dados falarem por si mesmos”. Henderson e Parmeter (2014) explicam que os estimadores paramétricos são considerados globais (usam todos os pontos), enquanto que os estimadores *kernels* não paramétricos usam uma amostra dos dados próximas a um ponto para ajustar a estimação, e a partir deste ajuste constrói-se um estimador global. A grande vantagem aqui é que isto permite focar nas peculiaridades locais relativas ao banco de dados utilizado, enquanto a função desconhecida (sem forma funcional preestabelecida) é estimada.

Mais especificamente, os autores esclarecem, de uma forma simplista, como este processo é feito. O valor médio da variável dependente é calculado em valores específicos de cada covariável.

Vershelde et al. (2012) demonstram que a média condicional ($E[Y|X = x]$) é estimada pelas médias de regressões localizadas ($E[Y|X \text{ próximo de } x] = g(x)$). Assim sendo, o emprego da regressão *kernel* não paramétrica para analisar a relação entre o tamanho das propriedades agrícolas e a produtividade terá estimativas independentes de formas funcionais, irá mostrar a forma sem restrições na estrutura de formação dos dados desta relação (suas não linearidades) e trará informações locais (se há diferenças na relação entre os tamanhos das propriedades) e globais (considerando todas as propriedades) ao mesmo tempo.

Contudo, como todos os métodos estatísticos, ele também apresenta desvantagens. Devido ao uso de uma amostra de dados próximas a um ponto para a estimação do modelo global, quanto maior a dimensão, maior a dificuldade para o estimador perceber a estrutura dos dados. Este fenômeno estatístico é chamado de “maldição da dimensionalidade”.

Henderson e Parmeter (2014) definem este fenômeno como sendo a dificuldade em detectar a estrutura do processo de geração de dados em altas dimensões sem excessivas suposições paramétricas. Em uma dimensão é fácil perceber a estrutura dos dados, visto que eles estarão dispostos em uma reta, sendo necessário o modelo somente preencher o espaço entre um ponto e outro (sendo com uma constante ou uma linha⁵) horizontalmente. Quando a dimensão aumenta para duas, faz-se necessário preencher o espaço entre um ponto e outro, tanto horizontalmente quanto verticalmente, visto que eles não estarão mais dispersos em uma linha, e sim em um eixo cartesiano. A Figura 1 busca melhor ilustrar esta questão em uma e duas dimensões.

Figura 1. Representação da maldição da dimensionalidade em uma e duas dimensões

Fonte: Elaborada a partir de Henderson e Parmeter (2014).

Os autores ainda explicam a maldição da dimensionalidade como sendo a dificuldade em preencher o espaço entre os pontos, medido através do cálculo da distância Euclidiana entre os pontos mais distantes da amostra. Suponha-se que as variáveis sempre se encontram em $[0, 1]$ (independente da dimensão). Em uma dimensão o espaço terá 1 de comprimento. Em duas dimensões, terá $\sqrt{2}$, e generalizando em q dimensões terá $\sqrt{q}$. Se as variáveis se encontram entre $[-1, 1]$, em uma dimensão o espaço terá comprimento 2, em duas dimensões

⁵ Depende do estimador utilizado, que será melhor definido e explicado na parte da metodologia.

4, e em q dimensões terá 2^q . Em altas dimensões a dificuldade em preencher o espaço, devido ao seu tamanho, aumenta exponencialmente.

Há ainda outra diferença entre os métodos paramétrico e não paramétrico explicitada por eles. O primeiro oferece maior precisão com menos dados e independe da dimensionalidade. O segundo evita erros de especificação e a precisão aumenta com o tamanho da amostra, mas está sujeito a maldição da dimensionalidade. Este é o *trade-off* entre modelar parametricamente e não parametricamente.

Para tal trabalho, as principais vantagens do método não paramétrico é identificar a verdadeira estrutura dos dados com relação ao tamanho das propriedades agrícolas e a produtividade no Brasil, sem influência de formas funcionais; captar as não linearidades do modelo permitindo identificar relações locais dos dados ao mesmo tempo que gera um estimador global; e uma alta precisão devido ao tamanho da amostra. A maldição da dimensionalidade é a principal desvantagem, pois a mesma restringe a dimensão da função de produção utilizada para a análise.

Além disto, o uso da regressão não paramétrica para tal objetivo vem sendo empregado na literatura internacional. Através da literatura revisada, percebe-se que esse método ainda não foi aplicado para o Brasil. A seção seguinte traz os trabalhos que fazem o uso dele, seguido da realidade brasileira (estudos já feitos sobre o Brasil).

3.1 Uso da regressão não paramétrica para a análise da relação entre o tamanho das propriedades agrícolas e a produtividade no mundo

De acordo com a literatura pesquisada, Barret (1996) foi o primeiro a utilizar a regressão não paramétrica para analisar a relação entre o tamanho das propriedades agrícolas e a área para Madagascar. Ele utilizou o estimador Naradaya-Watson empregado com um *kernel* Epanechnikov para regredir o valor da produção por área em logaritmos pelo logaritmo da área; o logaritmo da área, em ares (medida de área que cem metros quadrados), por adultos equivalentes pelo logaritmo da renda per capita e pelo logaritmo da área per capita. Na primeira estimação a relação foi inversa enquanto que na segunda foi direta. Em uma nota de rodapé o autor credita a primeira aplicação deste método para uma análise empírica a Deaton (1989).

Deaton (1989) utilizou a análise não paramétrica para estudar a relação entre o preço do arroz e a distribuição de renda na Tailândia. Os valores foram transformados em

logaritmos antes de serem regredidos. Um estimador específico não é empregado, e sim um método mais simples. O formato da curva de densidade nos pontos é estimado através do número de pontos ao redor do ponto em questão. O gráfico é construído aplicando esta técnica em todos os pontos do eixo horizontal e contando o número de pontos que se enquadram nesta banda (quantidade de pontos considerados ao redor). Esta contagem é ponderada pelo número total de observações, o que define a densidade nesse ponto. A regressão, por sua vez, é feita de modo parecido. Em um dado valor da variável independente, é calculada a média ponderada dos valores da variável dependente próximas a este ponto. A ponderação é a mesma utilizada para construir a densidade (quanto mais pontos próximos, maior peso é dado a ele), mas são dimensionados através da estimativa da densidade neste ponto. O autor afirma que, com o aumento da amostra, este valor irá convergir para a esperança da média condicional da variável dependente dado o valor da variável independente.

Apesar de não ser utilizado para o mesmo propósito, o método empregado acima é o mais simples para a estimação *kernel* não paramétrico observado na literatura consultada. Através dele muitos estimadores foram desenvolvidos, inclusive os utilizados nesta dissertação. O uso da regressão *kernel* multivariada não paramétrica foi encontrada somente em trabalhos mais recentes. Henderson e Parmeter (2014) acreditam que o aumento da capacidade computacional, somada ao desenvolvimento de softwares (como o R) permitiu o maior uso destes métodos em análises empíricas.

O estimador utilizado por Barret (1996) também foi utilizado por Assunção e Braidó (2007) e Barret et al. (2010). Ambos os trabalhos utilizaram uma regressão uni variada para constatar a relação inversa e motivar a continuação de seus respectivos estudos. Os valores utilizados foram transformados em logaritmos, sendo que o primeiro utilizou o valor nominal das principais produções em rupias indianas por acre, e o segundo quilograma de arroz por are como variável dependente e a área cultivada como independente. A relação inversa foi constatada em ambos os casos, e os procedimentos já abordados na revisão de literatura foram feitos em seguida.

Verschelde et al. (2013) usou um banco de dados para o Burundi, que continha variáveis contínuas, discretas ordenadas e discreta desordenadas para estimar a relação abordada até agora. Foi estimada uma função de produção com o logaritmo do valor da produção por hectare como medida de produtividade, o logaritmo da área como principal variável explicativa e outras variáveis relativas a características das famílias e propriedades. Para tal foi usado um estimador linear local em detrimento do apresentado até agora, porque

ele tem melhores propriedades que serão apresentadas em seções posteriores. Foi utilizado três diferentes funções *kernels* para cada tipo de variável. Para dar peso às contínuas uma função *kernel* normal padrão foi utilizada; para ponderar as discretas desordenadas, Aitchison e Aitken (1976) *kernel*; e para as discretas ordenadas Wang e van Ryzin (1981) *kernel*. Em seguida, multiplicou-se todos os *kernels* para formar uma função *kernel* generalizada, que foi proposta por Racine e Li (2004). A relação inversa é observada e eles suportam a hipótese de que a relação inversa pode ser explicada por problemas na insegurança do preço dos alimentos e falhas nos mercados de terra e crédito.

Ali e Deininger (2015), através de uma regressão ponderada pelo *kernel* não paramétrica univariada, analisaram a relação entre o valor da produção da colheita por hectare com o tamanho das parcelas e propriedades e com o uso do trabalho por hectare, todos em logaritmos. O primeiro sofreu uma redução de 1.100 dólares por hectare para 100 dólares por hectare, e o segundo de 2.000 para 45 dias por hectare conforme o tamanho das propriedades e parcelas aumentaram. Eles também estimaram a eficiência técnica através de uma fronteira estocástica de produção junto com uma regressão polinomial local ponderada pelo *kernel* estimada através dela. Neste caso, o modelo falhou em suportar a relação inversa entre tamanho das propriedades e eficiência técnica.

Com a finalidade de analisar esta relação para o Nicarágua, Almeida (2016) fez uso de um painel para os anos de 1998, 2001 e 2005. A estratégia utilizada consistiu em comparar a análise paramétrica e não paramétrica. No caso paramétrico, foram feitas duas análises: Transversal dos dados (sem considerar painel) e para dados em painel, ambas com e sem o vetor de controle para as propriedades e famílias, com a intenção de gerar uma função de produção mais parcimoniosa e evitar a maldição da dimensionalidade ao utilizar a análise não paramétrica. Na não paramétrica, quatro diferentes estimadores foram empregados: Um designado especificamente para a análise de dados agrupados; um desenvolvido por Henderson et al. (2008) para efeitos fixos; outro para efeitos fixos apresentado por Li e Racine (2004); e por fim, um para efeitos aleatórios desenvolvido por Wang (2003). Para dar robustez aos resultados, diferentes especificações da função de produção foram usadas. Em um primeiro momento foi considerado o valor da produção como variável dependente e a área, os gastos com insumos e o trabalho como independentes. Depois, as variáveis (menos a área) foram divididas pela área e, mantendo-se a divisão pela área, a receita líquida da agricultura foi considerada a medida de produtividade; e por fim, uma especificação em que o trabalho familiar por área foi a variável dependente cujas independentes foram a área, gastos com

insumos por área, salários sombras dos trabalhadores familiares e o vetor com outras características relacionadas à propriedade e famílias. A relação inversa foi observada: Um aumento na área das propriedades leva um aumento pequeno na medida de produtividade.

3.2 A realidade brasileira

Não foram encontrados trabalhos que usem a abordagem de regressões *kernels* não paramétricas para realizar a análise da relação entre o tamanho das propriedades agrícolas e a produtividade no Brasil. Portanto, buscou-se apresentar os trabalhos que são relacionados ao tema.

Vicente et al. (2001) compararam a evolução da produção, do uso de insumos e da produtividade agrícola das Unidades da Federação para os anos de 1970, 1975, 1980, 1985 e 1995. Para isso, eles utilizaram dados do Censo Agropecuário, dados da Retrospectiva Agropecuária da Fundação Getúlio Vargas e do Instituto de Economia Agrícola. As ferramentas analíticas utilizadas para analisar a evolução do uso de insumos e da produtividade agregada e para explicar a influência das produtividades parciais na produtividade total foram, respectivamente, a teoria moderna dos números índices e o modelo de dados em painel com efeitos aleatórios estimados através do método dos mínimos quadrados generalizados. As principais conclusões dos autores foram que: A produção agrícola brasileira (setor de lavouras) cresceu em média 3,2% ao ano no período; o uso de fatores de produção atingiu o auge no ano de 1985, caindo em 1995, que pode ser explicado pela falta de cobertura deste último Censo Agropecuário; a produtividade parcial da terra e do trabalho cresceram 2,3% e 3,9% ao ano no período, respectivamente; a produtividade total dos fatores cresceu, em média, 2,8% ao ano, que está positivamente associada ao crescimento das produtividades parciais da terra e trabalho. Porém, os autores não objetivaram analisar se há alguma relação entre a produtividade agrícola e o tamanho das propriedades.

Por meio de um método não paramétrico conhecido como Análise Envoltória de Dados, Helfand e Levine (2004) estudaram os determinantes da eficiência técnica e a sua relação com o tamanho das propriedades agrícolas do Centro Oeste brasileiro. Foram usados dados do Censo Agropecuário do ano de 1995 e 1996 de todos os municípios desta região desagregados em 15 tamanhos de propriedades e 4 tipos de posse de terra (arrendatários, proprietários, compartilhados e ocupados). O valor líquido da produção agropecuária foi considerado a medida de produtividade, e os insumos foram divididos em cinco: Área

utilizada; trabalho; tratores; animais; e insumos comprados. Devido a problemas de confidencialidade, os dados não são apresentados por fazendas. Por isso, todas as fazendas de uma determinada faixa de área com o mesmo tipo de posse de terra, de cada município, foram agregadas e denominadas fazendas representativas. O logaritmo da ineficiência técnica foi regredido com o do tamanho das propriedades e estimado através do método dos Mínimos Quadrados Generalizados com efeitos fixos. Controles para as diferenças na composição da produção, acesso a instituições e bens públicos, e relacionados ao nível de tecnologia e uso de *inputs* foram feitos. Estes dois últimos são determinantes importantes para a eficiência desta região. A relação tamanho-eficiência tem o formato de U-invertido, e é inversa até as fazendas compreendidas entre 200-500 hectares quando somente os dois primeiros controles são considerados e até 1000-2000 hectares quando os outros são empregados, invertendo-se para as maiores faixas de área.

A produtividade total dos fatores (PTF) da agricultura e sua relação com investimentos na infraestrutura foram foco do trabalho de Mendes et al. (2009). Os investimentos em rodovia, pesquisa, telecomunicações, irrigação e energia elétrica tiveram o maior impacto na PTF em ordem decrescente. Isto explicita deficiências brasileiras que vêm a ser desafios para a produtividade agropecuária. O período analisado foi de 1985 a 2004 com um painel de dados estimado através dos Métodos dos Momentos Generalizados, em que se verificou o aumento da PTF do período de 1995 a 2004 com relação ao anterior. A implementação de instrumentos modernos de políticas públicas foi destaque para a contribuição no crescimento da PTF.

Nota-se que a PTF tem sido o foco de trabalhos voltados para o Brasil nesta área. Vicente et al. (2001), explicam este fato como sendo uma medida de progresso tecnológico que evita a especificação errada da função de produção, o que também é evitado a partir da regressão *kernel* não paramétrica, como explicitado em seções anteriores. Nesta linha, autores como Gasques et al. (2011), Rada e Buccola (2012), Rada e Valdez (2012) buscaram explicar a evolução da PTF da agricultura, assim como o impacto das políticas públicas na mesma durante os anos. Algumas diferenças metodológicas e de políticas analisadas são observadas entre os trabalhos, as quais serão omitidas aqui porque não fazem relação com o tamanho das propriedades agrícolas, principal objetivo deste estudo.

A maior referência brasileira é o trabalho de Helfand et al. (2014) porque é feita uma abordagem sobre a relação entre a produtividade total dos fatores e o tamanho das propriedades agrícolas brasileira. Os dados utilizados foram do Censo Agropecuário para os

anos de 1985 a 2006, divididos em 5 classes de tamanhos das propriedades. Para esquivar dos problemas de confidencialidade, as observações também foram agregadas em fazendas representativas. Índices para a produtividade e insumos (terra, estoque de capital, trabalho familiar e insumos comprados) foram criados. Uma análise descritiva detalhada sobre o aumento do uso de insumos e produtividade durante os anos foi feita. A análise da relação é feita pelo método da fronteira estocástica cuja função de produção empregada foi da forma translog, estimada considerando efeitos fixos. Os resultados foram apresentados em 3 tópicos, mais uma checagem de robustez, sendo eles: Crescimento da produtividade nacional e regional; crescimento da produtividade nacional por classe de área; crescimento da produtividade por regiões por classe de área. A média de crescimento anual da PTF foi de 1,74% para o Brasil, sendo 5,0% de mudança técnica e -3,6% de mudança de eficiência técnica. A única região cuja mudança técnica foi negativa e a mudança de eficiência técnica foi positiva foi o Centro Oeste. Os resultados de maior interesse são os relativos aos tamanhos das propriedades. Foi observada uma relação em formato de U, a qual as menores (0-5 hectares) e as maiores (500 hectares e mais) propriedades tiveram maior crescimento da PTF e as medianas os menores. Os resultados variaram bastante de acordo com as regiões, o que explicita a heterogeneidade da agropecuária brasileira.

Nos estudos sobre produtividade agrícola no Brasil, a maior parte visa analisar a evolução dela durante os anos. Helfand e Levine (2004) e Helfand et al. (2015) são os únicos autores, dos trabalhos analisados, que a relacionaram com o tamanho das propriedades. Além disso, é utilizada uma abordagem diferente, com enfoque na produtividade total dos fatores da agricultura e o emprego do método da fronteira estocástica. A principal diferença, fora as variáveis utilizadas, é a abordagem não paramétrica. Esta irá desconsiderar qualquer influência da função de produção translog considerada.

4 MATERIAIS E MÉTODOS

A estratégia a ser seguida é estimar uma função de produção Cobb-Douglas através de um modelo de regressão linear múltiplo paramétrico (método dos Mínimos Quadrados Ordinários) e modelo de regressão múltiplo *kernel* não paramétrico (através de 3 diferentes estimadores). É preferida a função Cobb-Douglas por ela ser mais parcimoniosa e evitar problemas com a dimensionalidade já discutidos. Na maioria dos trabalhos analisados a estimação não paramétrica faz uso dos valores em logaritmos. Ambos os resultados serão comparados.

Os resultados paramétricos foram estimados através do comando *regress* do *software Stata 12*. O SIDRA (sistema de recuperação de dados do IBGE) concedeu os dados em planilhas eletrônicas do *Microsoft Excel*. Aproveitou-se deste formato e tratou-se os dados inicialmente nele. Os não paramétricos, por sua vez, foram estimados através do *software R*. Os comandos utilizados foram desenvolvidos por Henderson e Parmeter (2014) e apresentados no livro “*Applied Nonparametric Econometrics*” e estão disponíveis no site www.the-smooth-operators.com.

4.1 A função de produção Cobb-Douglas e a relação entre a produtividade e o tamanho das propriedades agrícolas

O modelo de regressão linear múltiplo estuda a relação entre a variável dependente e as variáveis independentes. A forma genérica de uma regressão linear, segundo Greene (2002) é:

$$y = f(x_1, x_2, x_3, \dots, x_k) + \varepsilon = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + \dots + x_k\beta_k + \varepsilon \quad (1)$$

em que:

$y = \text{variável dependente}$

$x_1, x_2, x_3, \dots, x_k = \text{variáveis independentes}$ ⁶

$\varepsilon = \text{termo de erro}$

⁶ Greene (2002) também as denomina como variáveis explicativas, covariáveis e regressores.

As variáveis a serem consideradas são especificadas pela teoria empregada no modelo. Neste caso, como uma função de produção será estimada, a variável dependente é a medida de produtividade adotada, enquanto que as independentes os insumos considerados relevantes. O objetivo é estimar os parâmetros desconhecidos $(\beta_1, \beta_2, \beta_3, \dots, \beta_k)$ através dos dados utilizados para validar as preposições teóricas analisadas.

Algumas assunções são preestabelecidas em um modelo de regressão múltiplo (GREENE, 2002):

- 1) Linearidade: linearidade da relação entre as variáveis dependentes e independentes, ou seja, a relação $y = f(x_1, x_2, x_3, \dots, x_k) + \varepsilon = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + \dots + x_k\beta_k + \varepsilon$ é linear;
- 2) Posto completo: não deve existir uma relação linear exata entre todas as variáveis independentes do modelo;
- 3) Exogeneidade das variáveis independentes: $E[\varepsilon_i | x_1, x_2, x_3, \dots, x_k] = 0$, ou seja, os valores esperados dos erros do modelo não são uma função das variáveis independentes observadas;
- 4) Homoscedasticidade e não auto correlação: cada erro, ε_i , tem uma variância finita, σ^2 , e não é correlacionado com nenhum outro erro;
- 5) Dados gerados exogeneamente: o processo de geração das variáveis independentes $(x_1, x_2, x_3, \dots, x_k)$ deve ser independentes da geração dos erros, ε_i ;
- 6) Distribuição normal dos erros: os erros, ε_i , devem ter uma distribuição normal.

Ulveling e Fletcher (1970) definem uma função Cobb-Douglas geral como:

$$Y = aA^\beta X_2^{\gamma_1} X_3^{\gamma_2} e^\varepsilon \quad (2)$$

Para satisfazer a assunção 1), é necessário transformar o modelo em uma equação loglinear (GREENE, 2002). Para isto, aplica-se o logaritmo natural nos dois lados da equação (2):

$$\ln Y = \ln a + \beta \ln A + \gamma_2 \ln X_2 + \gamma_3 \ln X_3 + \varepsilon \quad (3)$$

A equação (3) também é conhecida como elasticidade constante (ULVELING e FLETCHER, 1970; GREENE, 2002). A elasticidade de Y com respeito as mudanças de X é $\partial \ln Y / \partial \ln X_k$, que não varia com o X_k .

Desta maneira, para visualizar a relação entre a produtividade e a área, propõe-se uma forma simplificada da equação (3), como feito por Lamb (2003) e pode-se verificar na equação (4):

$$\ln Y = \alpha + \beta \ln A + \gamma_i \ln X_i + \varepsilon \quad (4)$$

Em que α, β, γ_i são os parâmetros a serem estimados e ε é o termo de erro. Segundo Heltberg (1998), se Y for a medida de produtividade e A a medida de área cultivada, β será a elasticidade da medida de produtividade com relação a área, então $\beta < 1$ a relação entre a produtividade e área das propriedades agrícolas é considerada inversa. Mas, se Y for a medida de produtividade por área a relação é considerada inversa se $\beta < 0$.

O método utilizado para estimar os parâmetros no modelo paramétrico, como já especificado, é o Mínimo Quadrado Ordinário. Ele consiste em minimizar a soma dos erros ao quadrado do modelo. Mais especificamente, ele é deduzido por Greene (2002) da seguinte maneira:

$$\min_{\beta_0} S(\beta_0) = \sum_{i=1}^n \varepsilon_{i0}^2 = \sum_{i=1}^n (y_i - \mathbf{x}_i \boldsymbol{\beta}_0)^2 \quad (5)$$

Expandindo esta relação:

$$S(\beta_0) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\beta}_0'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}_0 \quad (6)$$

A condição necessária para a minimização da soma dos quadrados dos erros é:

$$\frac{\partial S(\beta_0)}{\partial \beta_0} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}_0 = \mathbf{0} \quad (7)$$

Se a inversa da matriz $\mathbf{X}'\mathbf{X}$ existe, que provém da afirmação do posto completo, então a solução será:

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (8)$$

Para que esta solução minimize a soma dos quadrados dos erros, tem-se que:

$$\frac{\partial^2 S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = 2\mathbf{X}'\mathbf{X} \quad (9)$$

Deve ser uma matriz definida positiva. Define-se $q = \mathbf{e}'\mathbf{X}'\mathbf{X}\mathbf{e}$ para qualquer vetor $\mathbf{e}$ composto por não zeros:

$$q = \mathbf{v}'\mathbf{v} = \sum_{i=1}^n v_i^2 \quad (10)$$

onde: $\mathbf{v} = \mathbf{X}\mathbf{e}$

A menos que todos os elementos de $\mathbf{v}$ sejam 0, q é positiva. Se $\mathbf{v}$ for 0, então $\mathbf{v}$ será uma combinação linear das colunas de $\mathbf{X}$, o que viola o pressuposto de posto completo da matriz $\mathbf{X}$. O valor de q será positivo para qualquer valor não nulo de $\mathbf{e}$, de modo que $2\mathbf{X}'\mathbf{X}$ será positiva definida. Portanto, se o posto de $\mathbf{X}$ for completo, a solução de $\boldsymbol{\beta}$ é única e minimiza a soma dos quadrados dos resíduos.

4.2 O modelo de regressão *kernel* multivariado não paramétrico

O modelo não paramétrico empregado é semelhante ao paramétrico em sua especificação, a função de produção Cobb-Douglas (como já explicado, por ser mais parcimonioso) é utilizada em sua forma loglinear, como em todos os trabalhos analisados na literatura, como na equação (3).

No entanto, três estimadores foram utilizados para a estimação, o estimador linear constante local (LCLS), o estimador linear local (LLS) e o estimador polinomial local, o qual o mais conhecido, e também utilizado, é o estimador mínimo local quadrático (LQLS).

Todos os três serão definidos e comentados segundo o livro “*Applied Nonparametric Econometrics*” publicado por Henderson e Parmeter (2014). Para isso, usou-se definições e deduções feitas pelos autores nos capítulos 2 e 3, apenas comentadas aqui. No capítulo 2 é demonstrado como é feita a estimação da densidade no caso univariado por um método mais simples, avançando até o método *kernel*.

É introduzida a ideia da largura da janela, que determina a quantidade de observações usadas para construir a função densidade em um determinado ponto. A escolha dele é mais importante que a do *kernel*, porque ele é o *trade-off* entre o viés e a variância (quanto menor a largura da janela menor viés e maior a variância). A sua escolha ótima é feita através da minimização do erro assintótico integrado quadrático médio (AMISE). Menores larguras das janelas são desejáveis quanto maior o tamanho da amostra, pois maiores informações disponíveis permitem o estimador ser mais local. No entanto, este método depende do formato desconhecido da densidade subjacente da amostra. Para ser possível aplicá-lo, o formato desconhecido é substituído pelo formato de uma distribuição usada como referência. A família paramétrica da distribuição normal com média 0 e variância σ^2 é geralmente utilizada.

Silverman (1986) popularizou esta ideia. Com o uso de um *kernel* gaussiano, tem-se a regra geral de Silverman para largura da janela. Neste caso, a largura da janela ótima será dado pela fórmula:

$$h_{rg} = 1,06\hat{\sigma}n^{-1/5} \quad (11)$$

em que:

$\hat{\sigma}$ = desvio padrão da variável cuja densidade está sendo estimada

Este método é computacionalmente simples de calcular, mas, por outro lado, ele não tem fundamentos estatísticos formais quando o formato da densidade desvia do formato da densidade assumida. Quando a densidade não é normal a performance da largura da janela calculada por este método pode ser afetada (no sentido ótimo). Os métodos orientados pelos dados evitam estes problemas porque não dependem do formato da densidade. A validação cruzada de mínimos quadrados (*LSCV-Least Squares Cross Validation*) é o algoritmo orientado pelos dados mais populares para a seleção da largura da janela.

O objetivo da LSCV é minimizar a diferença entre o estimador da densidade e a própria densidade com relação a largura da janela. O apêndice técnico 2.13 do livro de Henderson e Parmeter (2014) demonstra que a forma mais comum em que ele é expresso é a abaixo:

$$LSCV_{(h)} = \frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n \bar{k} \left(\frac{x_i - x_j}{h} \right) - \frac{2}{n(n-1)h} \sum_{i=1}^n \sum_{j=1, j \neq i}^n k \left(\frac{x_i - x_j}{h} \right) \quad (12)$$

onde:

k = kernel

$\bar{k}$ = é chamado de convolution kernel (o mesmo apêndice discute como obtê-lo)

Embora esta ideia seja simples e eficiente, ela é computacionalmente dispendiosa, principalmente no caso multivariado (caso deste trabalho). Outros métodos orientados pelos dados existem, que serão abstridos aqui⁷.

No capítulo 3 os autores utilizam a intuição desenvolvida no capítulo anterior para apresentar os métodos multivariados não paramétricos (necessários para se desenvolver os estimadores). Um *kernel* multivariado é utilizado para este propósito em maiores dimensões.

Em suma, assim como no *kernel* univariado, o multivariado é dado como uma função de densidade de probabilidade multivariada que atribui pesos não negativos para pontos

⁷ O capítulo 2 do livro de Henderson e Parmeter (2014) traz e discute mais alguns destes métodos.

próximos via um vetor de largura da janela. A razão para se usar um vetor ao invés de uma única largura da janela é que as diferentes variáveis têm diferentes níveis de formato de densidade que oferecem uma configuração mais geral além de suas diferenças em escalas. Em trabalhos aplicados, é comum se utilizar o *kernel* produto, definido abaixo:

$$K_h(\mathbf{x}_i, \mathbf{x}) = \prod_{d=1}^q k\left(\frac{x_{id}-x_d}{h_d}\right) \quad (13)$$

onde:

$k(.)$ = kernel

Desta maneira, estimador de densidade *kernel* multivariado, se o *kernel* produto é adotado, é construído como:

$$\hat{f}(\mathbf{x}) = (n|\mathbf{h}|)^{-1} \sum_{i=1}^n K_h(\mathbf{x}_i, \mathbf{x}) \quad (14)$$

em que:

$|\mathbf{h}| = h_1 h_2 \dots h_q$ = largura da janela referente a cada variável

No caso univariado, o estimador é construído pela alavancagem das informações dos pontos próximos, com a proximidade sendo indicada pela largura da janela. No multivariado a ponderação é feita em múltiplas dimensões (mais especificamente, q-dimensões). Aqui se dá o problema da “maldição da dimensionalidade” já discutido. Quanto mais dimensões, mais difícil identificar a estrutura do processo de geração dos dados.

O processo de escolha das larguras das janelas é semelhante ao já discutido. A diferença é que ao invés de 1 variável (h e x) existirão q variáveis. A regra geral da largura da janela e o LSCV para o caso multivariado são representados abaixo, respectivamente:

$$h_d^{rg} = c(q) \widehat{\sigma_d} n^{-\frac{1}{4+q}} \quad (15)$$

onde:

$c(q)$ = é o fator escalar definido de acordo com o número (q) de variáveis⁸.

⁸ A tabela 3.3 do livro traz os fatores de q=2 até q=10 para 4 diferentes *kernels*

$$LSCV(\mathbf{h}) = \frac{1}{(n-1)|\mathbf{h}|} \sum_{i=1}^n \sum_{j=1}^n \bar{K}_h(\mathbf{x}_i, \mathbf{x}_j) - \frac{2}{n(n-1)|\mathbf{h}|} \sum_{i=1}^n \sum_{j=1, j \neq i}^n K_h(\mathbf{x}_j, \mathbf{x}_i) \quad (16)$$

em que:

$\bar{K}_h(\mathbf{x}_i, \mathbf{x}_j)$ = convolution kernel produto

Os estimadores *kernel* multivariados apresentados são importantes, mas não são o principal objeto de interesse, e sim a distribuição e a densidade condicional. Da densidade condicional pode-se investigar o primeiro momento (médias condicionais ou modelos de regressões), e os *quantis*, que são de maior interesse para os fins deste trabalho. O estimador da média condicional *kernel* vai combinar o estimador univariado com o multivariado para criar um estimador da densidade condicional.

Os estimadores das densidades conjuntas ($\hat{f}_{x,y}(\mathbf{x}, y)$) e marginais ($\hat{f}_x(\mathbf{x})$), para o caso de y uni variado (pode ser generalizado, mas não é o caso), é:

$$\hat{f}_{x,y}(\mathbf{x}, y) = (n|\mathbf{h}|h_y)^{-1} \sum_{i=1}^n K_h(\mathbf{x}_i, \mathbf{x}) k\left(\frac{y_i - y}{h_y}\right) \quad (17)$$

$$\hat{f}_x(\mathbf{x}) = (n|\mathbf{h}|)^{-1} \sum_{i=1}^n K_h(\mathbf{x}_i, \mathbf{x}) \quad (18)$$

sendo que:

h_y = parâmetro de suavização associado a y

$k(\cdot)$ = kernel uni variado

Estatisticamente, a densidade condicional de y dado $\mathbf{x}$ é definida como $f_{y|\mathbf{x}}(y|\mathbf{x}) = f_{x,y}(\mathbf{x}, y)/f_x(\mathbf{x})$. A densidade condicional estimada é:

$$\hat{f}_{y|\mathbf{x}}(y|\mathbf{x}) = \hat{f}_{x,y}(\mathbf{x}, y)/\hat{f}_x(\mathbf{x}) \quad (19)$$

Ela pode ser reescrita da seguinte maneira:

$$\hat{f}_{y|\mathbf{x}}(y|\mathbf{x}) = (nh_y)^{-1} \sum_{i=1}^n A_i(\mathbf{x}) k\left(\frac{y_i - y}{h_y}\right) \quad (20)$$

onde:

$$A_i(\mathbf{x}) = \frac{K_h(\mathbf{x}_i, \mathbf{x})}{n^{-1} \sum_{j=1}^n K_h(\mathbf{x}_j, \mathbf{x})} \quad (21)$$

A densidade condicional nesta forma revela que ela é construída como uma média local, com os pesos determinados pela densidade relativa do ponto $\mathbf{x}$ e calculando a média sobre o y suavizado. Ou seja, foi construído uma densidade incondicional de y e depois suavizando ela com o suporte da covariável, novamente utilizando a ideia de localidade. Há duas fontes de médias locais ocorrendo: Uma na direção de y ; e a outra no domínio de $\mathbf{x}$.

Com respeito as larguras de janelas, os métodos orientados pelos dados são os mais indicados. O LSCV é o mais utilizado e a ideia é a mesma por trás da dos estimadores uni e multivariados para a densidade.

De um modo geral, em uma análise paramétrica, uma forma funcional $m(\cdot)$ é assumida conhecida até um número finito de parâmetros desconhecidos. Se é assumido que $m(\mathbf{x}_i) = \alpha + \mathbf{x}_i\beta$, o modelo pode ser estimado pelo Mínimo Quadrado Ordinário. Mas se a especificação paramétrica da média condicional desvia de $m(\mathbf{x}_i) = \alpha + \mathbf{x}_i\beta$, o MQO será viesado e inconsistente. Independentemente da forma exata de $m(\mathbf{x}_i)$ podemos interpretar isso como a média condicional de y , $E(y|\mathbf{x})$. Logo, Henderson e Parmeter (2014) desenvolvem um estimador não paramétrico para essa quantidade através dos resultados do estimador da densidade condicional já discutido. A partir do estimador LCLS os outros serão deduzidos.

4.3 O estimador constante local (LCLS-*local constant linear squares*)

Henderson e Parmeter (2014) trazem três diferentes derivações do LCLS. A primeira é a partir da derivação dos estimadores de densidade; o segundo utiliza uma via intuitiva; e o terceiro, utiliza a mesma ideia do MQO paramétrico. No presente trabalho atentou-se para o terceiro, pois este já foi abordado neste texto na parte paramétrica.

Pode-se pensar em estimar a função desconhecida ($m(\mathbf{x}_i)$) minimizando o quadrado da distância ponderada entre a função e y . Isto nada mais é que uma média ponderada em que os pesos variam de acordo com $\mathbf{x}$.

Do MQO:

$$\min_{\alpha, \beta} \sum_{i=1}^n (y_i - \alpha - \mathbf{x}_i\beta)^2 \quad (22)$$

Pela condição de primeira ordem, iguala-se a 0 e obtêm-se a inclinação e o intercepto dos estimadores. Se, ao invés de $\alpha + \mathbf{x}_i\beta$, considerar $m(\mathbf{x}_i)$ e introduzir os pesos *kernels*:

$$\min_a \sum_{i=1}^n [y_i - m(\mathbf{x})]^2 K_h(\mathbf{x}_i, \mathbf{x}) \quad (23)$$

Da condição de primeira ordem:

$$-2 \sum_{i=1}^n (y_i - a) K_h(\mathbf{x}_i, \mathbf{x}) = 0 \quad (24)$$

Solucionando:

$$a = \hat{m}(\mathbf{x}) = \frac{\sum_{i=1}^n K_h(\mathbf{x}_i, \mathbf{x}) y_i}{\sum_{i=1}^n K_h(\mathbf{x}_i, \mathbf{x})} \quad (25)$$

Na essência, regrediu-se uma constante, localmente, em y para determinar a função no ponto. Na forma matricial esse estimador pode ser escrito:

$$\hat{m}(\mathbf{x}) = (\iota' K(\mathbf{x}) \iota)^{-1} \iota' K(\mathbf{x}) y \quad (26)$$

Onde ι é um vetor $n \times 1$ de 1 e:

$$K(\mathbf{x}) = \begin{bmatrix} K_h(\mathbf{x}_1, \mathbf{x}) & \dots & 0 \\ 0 & \ddots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & K_h(\mathbf{x}_n, \mathbf{x}) \end{bmatrix} \quad (27)$$

A largura da janela ótimo é selecionado pelo LSCV. Neste caso, o LSCV pode ser definido como:

$$LSCV_{(h)} = \sum_{i=1}^n [y_i - \hat{m}_{-i}(\mathbf{x}_i)]^2 \quad (28)$$

onde:

$$\hat{m}_{-i}(\mathbf{x}_i) = \frac{\sum_{j=1, j \neq i}^n y_j K_h(\mathbf{x}_j, \mathbf{x}_i)}{\sum_{j=1, j \neq i}^n K_h(\mathbf{x}_j, \mathbf{x}_i)} \quad (29)$$

Os autores destacam dois pontos importantes. Primeiro, $m(\cdot)$ é substituído pelo termo ao lado esquerdo da equação, y . Isto ocorre porque $m(\cdot)$ não é observado, mas sim uma versão corrompida ($y = m(x) + u$). Devido a assunção de que $E(u|x) = 0$, na média, $y_i = m(x_i)$. O segundo é que com a presença do $\hat{m}_{-i}(x_i)$, o algoritmo requer n^2 cálculos por avaliação da função de validação cruzada (LSCV). A LSCV deve ser rearranjada para que envolva apenas n , porque isto implica em um processo menos custoso computacionalmente.

Para dispensar $\hat{m}_{-i}(x_i)$, considera-se outra função objetivo de validação cruzada que pode ser escrita como:

$$\sum_{i=1}^n [y_i - \hat{m}(x_i)]^2 \pi(x_i) \quad (30)$$

Onde $\pi(\cdot)$ é uma função designada para aumentar com a rugosidade⁹ da função. O apêndice 5.6 do livro demonstra a forma desta função e foi omitida aqui. A intenção da largura da janela é a mesma da já apresentada: encontrar o valor mínimo que otimize o *trade-off* entre o viés e a variância.

Contudo, este estimador considera os pesos entre os pontos como uma constante, ou seja, estima-se a função com um ajuste constante. Isto pode causar problema visto que eles não são. Além disto, conforme a largura da janela aumenta, ela foca mais em aspectos globais que locais (em termos de ponderação).

4.4 Estimador linear local (LLLS-Local linear least squares)

Ao invés de uma constante, como o estimador anterior, o estimador linear local faz uso de uma linha local entre os pontos, o que, segundo Henderson e Parmeter (2014), provê características aplicadas e teóricas mais desejadas. Ele pode ser visto como uma expansão de Taylor no ponto x . Assim sendo, para a relação $y = m(x) + u$ tem-se os dados para $(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)$, e como tal, para cada ponto x_i , pode-se fazer uma aproximação de Taylor para x . Naturalmente, é esperado que para pontos distantes de x a aproximação de $m(x)$ seja ruim. Mas, limitando a atenção localmente, via ponderação pelo *kernel*, pode-se mitigar a influência das aproximações ruins.

⁹ Henderson e Parmeter (2014) demonstram no apêndice técnico do livro *Applied Nonparametric Econometrics*

que $\pi(x_i) = \left(\frac{\sum_{j=1}^n K_h(x_j, x_i)}{\sum_{j=1, j \neq i}^n K_h(x_j, x_i)} \right)^2$.

Considera-se a expansão de Taylor sobre $\mathbf{x}$ para as i observações:

$$y_i = m(\mathbf{x}_i) + u_i \approx m(\mathbf{x}) + (\mathbf{x}_i - \mathbf{x})\beta(\mathbf{x}) + u_i \quad (31)$$

Onde $(\mathbf{x}_i - \mathbf{x})$ é um vetor de dimensão $1 \times q$ e $\beta(\mathbf{x})$ é um vetor (coluna) de gradientes de dimensão q . Ignorando os termos de ordem superiores e tratando $m(\mathbf{x})$ e $\beta(\mathbf{x})$ como parâmetros:

$$y_i = a + (\mathbf{x}_i - \mathbf{x})b + u_i = (1, (\mathbf{x}_i - \mathbf{x})) \begin{pmatrix} a \\ b \end{pmatrix} + u_i \quad (32)$$

O problema de minimização para o conjunto completo de n observações é:

$$\min_{a,b} \sum_{i=1}^n [y_i - a - (\mathbf{x}_i - \mathbf{x})b]^2 K_h(\mathbf{x}_i - \mathbf{x}) \quad (33)$$

Na forma matricial:

$$\min_{\delta} (y - \mathbf{X}\delta)' K(\mathbf{x}) (y - \mathbf{X}\delta) \quad (34)$$

Onde $\delta \equiv (a, b)'$, $\mathbf{X}$ é uma matriz de dimensão $n \times (q + 1)$ com a i -ésima linha sendo $(1, (\mathbf{x}_i - \mathbf{x}))$ e $K(\mathbf{x})$ é uma matriz diagonal de dimensão $n \times n$ com o i -ésimo elemento igual a $K_h(\mathbf{x}_i, \mathbf{x})$. Em outras palavras, é a generalização não paramétrica do método dos mínimos quadrados. Se $K(\mathbf{x})$ for uma matriz identidade, tem-se o estimador MQO. Se ele for a matriz de variância e covariância dos erros, tem-se o estimador dos mínimos quadrados generalizados. Ao minimizar a função objetivo com respeito a δ tem-se:

$$\hat{\delta}(\mathbf{x}) = \begin{pmatrix} \hat{m}(\mathbf{x}) \\ \hat{\beta}(\mathbf{x}) \end{pmatrix} = (\mathbf{X}' K(\mathbf{x}) \mathbf{X})^{-1} \mathbf{X}' K(\mathbf{x}) y \quad (35)$$

O LCLS utiliza uma constante para ponderar os pontos próximos, o LLLS faz uma expansão de Taylor de primeira ordem (uma linha) ao redor de $\mathbf{x}$ para obter uma melhor performance. Se, ao invés uma expansão de Taylor de primeira ordem, utilizar uma de ordens maiores, tem-se o estimador polinomial local. Henderson e Parmeter (2014) afirmam que o quadrático (LQLS) é o mais popular.

4.5 Estimador quadrático local (LQLS- *Local quadratic least squares*)

Como já introduzido, o estimador polinomial local utiliza expansões de Taylor ao redor de $\mathbf{x}$ maiores que a primeira. No entanto, a escolha de quantas expansões serão feitas é muito importante. Mais expansões reduzem o viés, mas aumentam a variância (HENDERSON e PARMETER, 2014). Isto é causado pelo aumento do número de parâmetros que deve ser estimado.

Este aumento também nos leva a já conhecida “maldição da dimensionalidade”. Sendo p a ordem de expansão do estimador polinomial e q o número de covariáveis, o número de parâmetros a serem estimados é $\binom{p+q}{p} = \frac{(p+q)!}{p!q!}$. Há um *trade-off* aqui, quanto maior a ordem, melhor a estimação do formato desconhecido da função, mas mais custoso operacionalmente ele será. Utilizou-se o LQLS, pois através dele as informações necessárias para este estudo são satisfeitas. A expansão de Taylor de segundo grau gera a equação:

$$y_i \approx m(\mathbf{x}) + (\mathbf{x}_i - \mathbf{x}) \frac{\partial m(\mathbf{x})}{\partial \mathbf{x}} + \frac{1}{2} \text{vec}[(\mathbf{x}_i - \mathbf{x})(\mathbf{x}_i - \mathbf{x})']' \text{vec} \left[\frac{\partial^2 m(\mathbf{x})}{\partial \mathbf{x}' \partial \mathbf{x}} \right] + u_i \quad (36)$$

Em que $\text{vec}(A)$ representa a vetorização da matriz A (coluna). $\partial^2 m(\mathbf{x}) / \partial \mathbf{x}' \partial \mathbf{x}$ inclui a segunda derivada e também as derivadas parciais, porque $\frac{\partial^2 m(\mathbf{x})}{\partial x_i \partial x_j} = \frac{\partial^2 m(\mathbf{x})}{\partial x_j \partial x_i}$. O problema de minimização para n observações é:

$$\min_{a_0, a_1, a_2} \sum_{i=1}^n [y_i - a_0 - (\mathbf{x}_i - \mathbf{x})a_1 - \text{vec}[(\mathbf{x}_i - \mathbf{x})(\mathbf{x}_i - \mathbf{x})']' a_2]^2 K_h(\mathbf{x}_i, \mathbf{x}) \quad (37)$$

O estimador de dimensão $((q^2 + 1) \times 1)$ é:

$$\hat{\delta}(\mathbf{x}) = \left(\hat{m}(\mathbf{x}), \frac{\partial \hat{m}(\mathbf{x})}{\partial \mathbf{x}}, \text{vec} \left[\frac{\partial^2 \hat{m}(\mathbf{x})}{\partial \mathbf{x}' \partial \mathbf{x}} \right] \right)' = [\mathbf{X}' K(\mathbf{x}) \mathbf{X}]^{-1} \mathbf{X}' K(\mathbf{x}) \mathbf{y} \quad (38)$$

Agora a inésima linha de $\mathbf{X}$ é definida como $\mathbf{X}_i \equiv (1, (\mathbf{x}_i - \mathbf{x}), \text{vec}[(\mathbf{x}_i - \mathbf{x})(\mathbf{x}_i - \mathbf{x})']')$. Percebe-se facilmente que o número de estimações locais aumenta rapidamente com o aumento do q . Com isso, define-se os 3 estimadores que serão utilizados para realizar a análise não paramétrica.

Contudo, não há um modo padrão de apresentação das estimativas feitas como as tabelas de resultados paramétricas com os coeficientes e respectivos erros padrões. Há diversos métodos que podem ser utilizados para apresentar os resultados não paramétricos das estimativas das regressões. Utilizou-se a representação gráfica da relação das médias parciais da média condicional entre a variável independente e uma das covariáveis, mantendo as outras em fixas em seu valor médio (VERSCHELDE et al., 2013; HENDERSON; PARMETER, 2014); e em tabela com as elasticidades para os *quantis* 0,25, 0,50, 0,75 (HENDERSON; PARMETER, 2014).

As elasticidades são calculadas segundo a fórmula:

$$elasticidade(X_i) = \frac{\hat{\beta}(X_i)X_i}{\hat{Y}_i} \quad (39)$$

onde:

$\hat{\beta}(X_i)$ = gradiente da média condicional da covariável X_i

$\hat{Y}_i$ = é o valor estimado de Y

Sob a assunção de retornos constantes de escala, as elasticidades podem ser interpretadas como a porcentagem base de um dado insumo usado no processo de produção. Henderson e Parmeter (2014) enfatizam que ela pode ser interpretada desta maneira somente sob competição perfeita e retornos constantes de escala

4.6 Base de dados e estatística descritiva

Os dados utilizados foram coletados do Censo Agropecuário do ano de 2006, realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE) e recuperados através do Sistema IBGE de Recuperação Automática de Dados (SIDRA). O censo foi realizado no ano de 2007 através da aplicação de questionários eletrônicos em cada um dos estabelecimentos agropecuários visitados pelo IBGE, de modo a coletar informações sobre as atividades agropecuárias desenvolvidas. A partir destas informações foi criado o banco de dados utilizado neste trabalho.

Por fins de confidencialidade, o IBGE não disponibiliza os dados por estabelecimento, mas os agrega por município e faixa de área. Cada município tem seus estabelecimentos

agregados de acordo com a área, sendo dezessete faixas de área. As faixas de área que não tem estabelecimentos produtivos são excluídas da análise. Se não há estabelecimentos produtivos não há observação do fenômeno naquela faixa de área, daquele município. Cada faixa de área, por sua vez, consiste na soma dos dados coletados de todos os estabelecimentos referentes a ela. Por exemplo: O valor da produção da faixa de área X do município Y representa a soma do valor da produção de todos os estabelecimentos agropecuários de Y cuja área pertença a X.

Cada observação agregada considera homogêneo os estabelecimentos localizados no mesmo município, e dentro do mesmo conjunto de extensão territorial. Helfand et al. (2015) as denominam como “fazendas representativas”. Estas representam o comportamento de determinado grupo de estabelecimentos que partilham a mesma localização e mesmo tamanho. A análise econométrica explorou a variação entre elas. Porém, devido a agregação, não foi possível analisar a variação dentre os estabelecimentos que estão contidos em uma mesma fazenda representativa.

Assim sendo, o tamanho da amostra é potencialmente aumentado de 5.538 observações (relativo ao número de municípios coletados pelo IBGE) para 94.146 (número de municípios multiplicado pelo número de faixas de área). Porém, esta não será a amostra final para a análise econométrica. Algumas observações serão excluídas, as quais serão demonstradas e justificadas a seguir.

4.7 Variáveis utilizadas no modelo econométrico

As variáveis utilizadas na análise do presente trabalho são as mesmas para ambos os modelos paramétricos e não paramétricos, a fins de comparação entre eles. Em suma, as variáveis são: A medida de produtividade e os insumos utilizados (área produtiva, capital e trabalho). Eles são descritos detalhadamente a seguir.

O valor da produção (em mil reais) por hectare é utilizado como medida da produtividade da área. O IBGE considera como sendo o valor da produção animal, vegetal e da agroindústria. A produção animal é subdividida em grande, médio e pequeno porte, além das aves. A vegetal em lavouras temporárias e permanentes, horticultura, floricultura, silvicultura e extração vegetal. Além disto, o valor da produção é dividido pela área produtiva das propriedades, mensurada em hectares, e transformados em logaritmos. As observações

cujo valor da produção é 0 foram excluídas para evitar vieses resultantes de propriedades não produtivas.

De acordo com o objetivo proposto, a principal variável independente de interesse é o logaritmo da área produtiva das propriedades, a qual é mensurada em hectares. O IBGE disponibiliza a área total e desagrega em 15 utilizações diferentes para a terra (lavouras permanentes, lavouras temporárias, área plantada com forrageiras de corte, área para cultivo de flores, pastagens naturais, pastagens plantadas, pastagens plantadas degradadas, pastagens em boas condições, matas e/ou florestas naturais destinadas à preservação permanente ou reserva legal e exclusiva as áreas de preservação permanente e as em sistemas agroflorestais, matas e/ou florestas plantas com essenciais florestais, sistemas agroflorestais, tanques e áreas de águas para exploração de aquicultura, construções e benfeitorias, terras degradadas e inaproveitáveis).

Foram excluídas 4 que não são produtivas (matas e/ou florestas destinadas a preservação permanente e reserva legal, matas e/ou florestas naturais exceto os sistemas agroflorestais, terras degradadas e inaproveitáveis para a agropecuária), e, portanto, não pertinentes a análise pois gera-se um viés na estimação. O coeficiente estimado desta variável menor que 0 indica uma relação inversa entre a área e a produtividade, caso ele seja positivo a relação entre a área e a produtividade será direta (CORNIA, 1981; HELTBERG, 1998; CARLETTTO et al., 2013). Todos os trabalhos analisados que utilizam a análise de regressão para estimar uma função de produção e estabelecer esta relação fazem desta variável a de maior interesse. Se a área produtiva for igual a 0, a divisão das variáveis pela área será incoerente matematicamente, portanto elas também são excluídas da análise.

O capital foi construído através do logaritmo da soma do valor dos bens, em mil reais, de cada estabelecimento no ano, os quais são: Bens imóveis (prédios, instalações e outras benfeitorias, terras, lavouras permanentes e matas plantadas) e outros bens (veículos, tratores, máquinas e implementos, animais para reprodução, criação e outros fins) divididos pela área em hectares.

A mão de obra é um insumo essencial, visto que sem ela ainda não há produção agropecuária. O IBGE divide em dois tipos o trabalho: O familiar, caracterizado pelo parentesco com o proprietário; e o contratado, ou seja, aqueles que não são familiares. Eles são apresentados no censo pelo número de trabalhadores por classes de dia trabalhados¹⁰.

¹⁰ Menos de 60 dias; de 60 a 180 dias; 180 e mais dias.

Porém, usa-se o logaritmo das horas de trabalho por hectare de modo a incluir os empregados temporários na análise econométrica. Para isto considera-se 60, 120 e 180 dias trabalhados por ano e que a jornada de trabalho seja de 8 horas diárias. Multiplica-se os dias trabalhados no ano por 8 (horas) e pelo respectivo número de trabalhadores. Os dois tipos de trabalho são somados e tem-se o trabalho total em horas trabalhadas por ano.

4.8 Estatística descritiva das variáveis utilizadas no modelo econométrico

Além das observações excluídas descritas acima, excluiu-se as 1% menores e maiores propriedades por serem consideradas *outliers*. Eles são considerados um problema porque se comportam de uma maneira diferente do restante da amostra, podendo causar um viés na estimação. Em outras palavras, seus valores destoam em muito do restante dos dados, e a atipicidade de seus valores pode prejudicar a interpretação dos resultados. Soma-se 0,001 às variáveis, exceto na área produtiva. Este processo evita o descarte das variáveis cujos valores são 0 (BARRET et al., 2010). As estatísticas descritivas das variáveis seguem abaixo, separadas pela agregação, regiões e faixas de área (Tabelas 1, 2 e 3).

Desta maneira, quando a agregação municipal é considerada, o número de observações é de 5.193. Ao desagregar por faixa de área este número aumenta para 58.873. Este aumento é benéfico principalmente para a análise não paramétrica, que, como já explicado, gera um ganho de precisão na estimação. O ideal seria ter como observações cada propriedade agrícola (nível de microdados), porque as relações locais captadas pelos estimadores não paramétricos fariam referências a elas. Como isto não é possível, a variação captada por eles será, no máximo, dentre as faixas de áreas.

O primeiro aspecto que merece destaque é o elevado desvio padrão das variáveis. Ele pode ser entendido como uma alta disparidade entre seus valores máximos e mínimos. Isto reflete a heterogeneidade das propriedades agrícolas entre os municípios e faixas de área no Brasil.

A Tabela 1 apresenta as variáveis com o nível de agregação municipal para o Brasil e regiões. Após as exclusões, restaram 5.193 municípios. Cada município tem uma média¹¹ de 851 estabelecimentos produtivos, que por sua vez, tem 43,9 hectares. Somados eles ocupam 37.873 hectares e produzem 1.081 reais por hectare. Eles contam com 7.220 reais de capital e

¹¹ Todos os valores são com base na média das variáveis.

23,3 horas de trabalho por hectare. Vale ressaltar que os valores monetários são nominais do ano de 2006.

Tabela 1. Estatística descritiva por regiões considerando municípios como observações agregadas.

Variável	Obs.	Média	Desv. Pad.	Mín.	Máximo
Brasil					
Nº Estabelecimentos	5.193	851,6	911,2	6	9.565
VP/ha (mil reais)	5.193	1,081	1,215	0,0488	9,721
Área Prod. (ha)	5.193	37.387	49.286	731	381.855
Capital/ha (mil reais)	5.193	7,220	5,798	0,577	36,37
Trab. Tot./ha (horas)	5.193	23,30	21,08	1,156	147,8
Norte					
Nº Estabelecimentos	394	947,5	1.039	21	9490
VP/ha (mil reais)	394	0,467	0,748	0,0494	9,721
Área Prod. (ha)	394	65.977	60.648	793	346.378
Capital/ha (mil reais)	394	3,017	2,219	0,668	24,34
Trab. Tot./ha (horas)	394	17,26	21,46	1,169	147,8
Nordeste					
Nº Estabelecimentos	1.677	1.267	1.165	9	8.515
VP/ha (mil reais)	1.677	0,676	0,950	0,0488	9,439
Área Prod. (ha)	1.677	28.782	32.055	787	330.937
Capital/ha (mil reais)	1.677	3,009	2,644	0,577	27,04
Trab. Tot./ha (horas)	1.677	34,03	25,32	1,642	147,8
Centro Oeste					
Nº Estabelecimentos	400	558,7	479,0	26	3.845
VP/ha (mil reais)	400	0,480	0,458	0,0529	4,570
Área Prod. (ha)	400	109.759	93.203	4.075	381.855
Capital/ha (mil reais)	400	5,297	2,721	0,955	25,59
Trab. Tot./ha (horas)	400	5,095	4,592	1,156	51,93
Sudeste					
Nº Estabelecimentos	1.570	497,3	523,2	6	6.250
VP/ha (mil reais)	1.570	1,259	1,216	0,0505	9,471
Área Prod. (ha)	1.570	27.625	31.602	731	272.401
Capital/ha (mil reais)	1.570	9,553	5,595	1,111	36,26
Trab. Tot./ha (ha)	1.570	17,58	15,51	1,300	128,2
Sul					
Nº Estabelecimentos	1.151	798,0	699,6	7	9.565
VP/ha (mil reais)	1.151	1,846	1,396	0,0506	9,553
Área Prod. (ha)	1.151	28.311	37.597	784	358.957
Capital/ha (mil reais)	1.151	12,27	5,324	1,324	36,37
Trab. Tot./ha (horas)	1.151	23,89	15,42	1,529	109,3

Fonte: Elaborada a partir de dados do Censo Agropecuário de 2006 (IBGE, 2017).

Na região Norte, 394 municípios foram analisados. Observa-se um número superior de estabelecimentos¹² assim como de área produtiva. A média de área por estabelecimento é de 69,7 hectares, quase 60% superior. No entanto, eles produzem menos, possuem menos capital e empregam menos trabalho por hectare. Os dois primeiros representam menos que a metade da média nacional, enquanto o último aproximadamente um terço.

O Nordeste é a região com maior número de municípios e estabelecimentos por município. Sua área produtiva é inferior, o que indica a presença de pequenas propriedades. Os estabelecimentos têm 22,7 hectares, aproximadamente metade do que os brasileiros. O valor da produção e o capital por hectare também são inferiores. Com relação ao trabalho, mais horas são usadas por hectare, ou seja, estas propriedades são mais intensivas em trabalho que a média nacional.

Com a maior área produtiva, os municípios do Centro Oeste possuem 109.759 hectares produtivos, 3 vezes a média nacional, sendo que cada estabelecimento tem uma média de 196 hectares. O valor da produção somente é superior aos nortistas, e também não atinge a metade da média nacional por hectare. O capital e o trabalho são inferiores, sendo que o segundo é o menor de todas as regiões.

Ao contrário do Centro Oeste, os municípios do Sudeste somam a menor área produtiva. O valor da produção e o capital por hectare são superiores, na medida que utilizam menos trabalho por hectare.

O Sul, por sua vez, é a região que apresenta maior valor de produção e capital por hectare. Ambos são aproximadamente 70% maiores. A área produtiva é inferior e próxima da do Sudeste. As horas de trabalho por hectare são equivalentes à média nacional.

A Tabela 2 traz as estatísticas descritivas por região considerando as fazendas representativas como observações agregadas. A diferença com relação a Tabela 1 é que agora as médias não são com relação somente aos municípios, e sim com relação a cada faixa de área de cada município. As diferenças de valores ocorrem porque faixas de área de determinados municípios que não possuem propriedades são excluídas. Assim sendo, os resultados presentes nela são com relação a cada fazenda representativa e não mais com os municípios.

¹² As comparações são feitas com base no Brasil.

Tabela 2. Estatística descritiva por região considerando as fazendas representativas como observações agregadas

Variáveis	Observações	Média	Desvio Padrão	Mínimo	Máximo
Brasil					
VP/ha (mil reais)	58.873	2,45	5,209	0,023	65,56
Área Prod. (ha)	58.873	2.462	4.795	2	43.127
Capital/ha (mil reais)	58.873	13,93	20,08	0,602	183,3
Trab Tot/ha (horas)	58.873	167,1	307,3	0,72	2.595
Norte					
VP/ha (mil reais)	4.365	2,321	5,811	0,023	63,33
Área Prod. (ha)	4.365	4.330	7.311	2	42.962
Capital/ha (mil reais)	4.365	7,094	13,08	0,602	174
Trab Tot/ha (horas)	4.365	176,7	350,5	0,722	2.580
Nordeste					
VP/ha (mil reais)	20.511	1,821	4,295	0,023	64,14
Área Prod. (ha)	20.511	1.899	3.551	2	43.124
Capital/ha (mil reais)	20.511	5,891	9,874	0,602	181,4
Trab Tot/ha (horas)	20.511	222,2	368,4	0,726	2.595
Centro Oeste					
VP/ha (mil reais)	4.567	1,424	3,851	0,0233	64,58
Área Prod. (ha)	4.567	5.335	8.457	2	43.127
Capital/ha (mil reais)	4.567	11,76	17,6	0,664	178,7
Trab Tot/ha (horas)	4.567	86,55	181,1	0,723	2.440
Sudeste					
VP/ha (mil reais)	16.862	2,759	5,688	0,024	65,33
Área Prod. (ha)	16.862	2.110	3.936	2	41.752
Capital/ha (mil reais)	16.862	19,38	23,27	0,627	181,8
Trab Tot/ha (horas)	16.862	130,7	254,3	0,72	2.590
Sul					
VP/ha (mil reais)	12.557	3,48	5,855	0,0245	65,56
Área Prod. (ha)	12.557	2.158	4.100	2	42.883
Capital/ha (mil reais)	12.557	22,88	24,27	0,614	183,3
Trab Tot/ha (horas)	12.557	152,2	264,9	0,722	2586

Fonte: Elaborada a partir dos dados do Censo Agropecuário de 2006. (IBGE, 2017).

As diferenças de magnitudes ocorreram para todas as variáveis. No entanto, a região Norte foi que os resultados mais se diferenciaram. O seu valor de produção por hectare se aproximou da média nacional e ultrapassou o do Nordeste e do Centro Oeste. O trabalho também sofreu um aumento substancial.

Vale ressaltar que nem todas as regiões possuem propriedades compreendidas em todas as faixas de área. Quanto maior a disparidade de tamanho entre as propriedades, mais faixas de área elas abrangem. Como cada faixa de área de um determinado município é uma observação, quanto mais faixas de área mais observações.

Devido a isto, a Tabela 2 não é um instrumento com uma boa qualidade para se comparar as variáveis dentre as regiões. Porém ela é de suma importância porque esta é a maior desagregação possível dos dados e será utilizado na análise de regressão.

Ao se observar a Tabela 3, que apresenta as estatísticas descritivas das fazendas representativas por faixa de área, é possível perceber melhor esta diferença. Há uma grande variedade nas observações entre as faixas de área. Isso não é um problema para a análise econométrica porque ela considera cada observação como se fosse uma propriedade agrícola.

Em um primeiro momento, percebe-se que os desvios padrões continuam maiores que a média, mantendo-se a indicação da heterogeneidade entre as propriedades agrícolas. Esta diferença é reduzida nas maiores faixas de área. O número de observações em cada faixa de área também chama a atenção.

As faixas de áreas compreendidas entre 5 a 100 hectares apresentam o maior número de observações, ou seja, elas são mais comuns nos municípios brasileiros. Elas ultrapassam o número de observações agregadas municipalmente. Isto ocorre por causa das exclusões já comentadas. Se elas ultrapassam o número de municípios agregados, significa que existem municípios que só contém propriedades destes tamanhos. Logo, os valores das variáveis para esses municípios são abruptamente diferentes e excluídos.

As propriedades produtivas compreendidas entre 5 e 50 hectares são as mais abundantes no Brasil. As menos abundantes são os dois extremos: As com mais de 2.500 hectares e entre 0 e 0,1 hectares. As fazendas representativas com 2.500 hectares ou mais têm, em média, 5,5 propriedades com 19.943 hectares. As de 0 a 0,1 hectares tem 31,6 estabelecimentos e que somam 3,2 hectares. Percebe-se também, que os tamanhos médios de propriedades são mais comuns no Brasil.

Tabela 3. Estatística descritiva por faixa de área

(Continua)

Variável	Obs.	Média	Desv. Pad.	Mín.	Máx.
Mais de 0 a menos de 0,1 ha					
Nº Estabelecimentos	158	31,6	21,61	12	63
VP/ha (mil reais)	5	5,053	2,075	2,5	7,5
Área Prod. (ha)	5	3,2	1,304	2	5
Capital/ha (mil reais)	5	31,15	15,55	6,4	45,5
Trab. Tot./ha (horas)	5	2.378	159,5	2.175	2.595
De 0,1 a menos de 0,2 ha					
Nº Estabelecimentos	18.718	46,68	59	3	458
VP/ha (mil reais)	401	9,003	10,83	0,25	59,5

(continuação)

Área Prod. (ha)	401	7,908	9,768	2	86
Capital/ha (mil reais)	401	44,11	41,2	0,8	178,5
Trab. Tot./ha (horas)	401	1.961	424,6	460	2.592
De 0,2 a menos de 0,5 ha					
Nº Estabelecimentos	137.815	69,01	131,3	3	2.334
VP/ha (mil reais)	1.997	6,557	9,414	0,0526	65,4
Área Prod. (ha)	1.997	25,35	48,38	2	852
Capital/ha (mil reais)	1.997	37,95	42,31	0,667	183,3
Trab. Tot./ha (horas)	1.997	1.134	407,3	270	2.555
De 0,5 a menos de 1 ha					
Nº Estabelecimentos	263.160	76,99	155,4	3	1.931
VP/ha (mil reais)	3.418	5,75	8,826	0,0862	65,33
Área Prod. (ha)	3.418	56,49	117,3	2	1426
Capital/ha (mil reais)	3.418	33,9	37,22	0,625	181,8
Trab. Tot./ha (horas)	3.418	622,9	244,3	120	2.456
De 1 a menos de 2 ha					
Nº Estabelecimentos	401.994	90,36	163,4	3	1.994
VP/ha (mil reais)	4.449	4,923	7,739	0,0256	65,56
Área Prod. (ha)	4.449	122	219,9	2	2311
Capital/ha (mil reais)	4.449	25,19	27,63	0,65	181,4
Trab. Tot./ha (horas)	4.449	337	137,3	70	2511
De 2 a menos de 3 ha					
Nº Estabelecimentos	285.829	60,72	92,56	3	1.477
VP/ha (mil reais)	4.707	3,549	5,984	0,0333	65,2
Área Prod. (ha)	4.707	142,7	207,4	3	3.006
Capital/ha (mil reais)	4.707	18,29	19,47	0,616	177,1
Trab. Tot./ha (horas)	4.707	202,1	95,83	40,87	2.237
De 3 a menos de 4 ha					
Nº Estabelecimentos	227.820	49,55	70,75	3	860
VP/ha (mil reais)	4.598	3,137	5,483	0,025	64,21
Área Prod. (ha)	4.598	166,2	232,8	8	2958
Capital/ha (mil reais)	4.598	15,85	16,9	0,654	178,2
Trab. Tot./ha (horas)	4.598	148,8	79,43	20	2.486
De 4 a menos de 5 ha					
Nº Estabelecimentos	188.973	39,56	51,61	3	837
VP/ha (mil reais)	4.777	2,568	4,64	0,0245	62,64
Área Prod. (ha)	4.777	180,8	229,9	8	3.481
Capital/ha (mil reais)	4.777	13,54	14,07	0,614	181
Trab. Tot./ha (horas)	4.777	114	71,45	18	2.566
De 5 a menos de 10 ha					
Nº Estabelecimentos	563.869	106,7	131,8	3	1.855
VP/ha (mil reais)	5.283	2,24	3,893	0,0262	59,88
Área Prod. (ha)	5.283	743,7	883,5	6	1.1631
Capital/ha (mil reais)	5.283	12,14	12,77	0,611	176,3
Trab. Tot./ha (horas)	5.283	78,98	54,86	15,93	1.624
De 10 a menos de 20 ha					

(conclusão)

Nº Estabelecimentos	660.636	123,1	145,9	3	2.830
VP/ha (mil reais)	5.366	1,689	3,196	0,0244	53,93
Área Prod. (ha)	5.366	1.593	1.790	19	33.973
Capital/ha (mil reais)	5.366	9,603	9,588	0,61	181,2
Trab. Tot./ha (horas)	5.366	45,77	46,24	10,46	2.231
De 20 a menos de 50 ha					
Nº Estabelecimentos	739.944	137,3	156,2	3	1.986
VP/ha (mil reais)	5.391	1,323	2,772	0,0235	54,63
Área Prod. (ha)	5.391	3.700	3.984	32	40.113
Capital/ha (mil reais)	5.391	7,984	7,29	0,602	113,3
Trab. Tot./ha (horas)	5.391	24,81	22,86	4,091	697,6
De 50 a menos de 100 ha					
Nº Estabelecimentos	332.460	63,52	84,99	3	1.124
VP/ha (mil reais)	5.234	0,969	2,224	0,0245	55,78
Área Prod. (ha)	5.234	3.644	4.230	60	42.463
Capital/ha (mil reais)	5.234	6,993	6,601	0,604	116,2
Trab. Tot./ha (horas)	5.234	13,52	14,67	1,875	373,5
De 100 a menos de 200 ha					
Nº Estabelecimentos	183.960	37,7	55,41	3	1.103
VP/ha (mil reais)	4.880	0,815	1,957	0,0235	38,21
Área Prod. (ha)	4.880	4.271	4.875	118	42.678
Capital/ha (mil reais)	4.880	6,347	6,094	0,602	127,3
Trab. Tot./ha (horas)	4.880	10,84	30,66	1,364	777,7
De 200 a menos de 500 ha					
Nº Estabelecimentos	113.570	26,41	28,84	3	433
VP/ha (mil reais)	4.301	0,673	1,675	0,023	44,07
Área Prod. (ha)	4.301	6.991	7.385	123	42.962
Capital/ha (mil reais)	4.301	5,856	6,4	0,603	136
Trab. Tot./ha (horas)	4.301	6,926	18,99	0,753	534,5
De 500 a menos de 1000 ha					
Nº Estabelecimentos	36.397	13,98	14,61	3	225
VP/ha (mil reais)	2.603	0,646	2,141	0,023	55,1
Área Prod. (ha)	2.603	8.018	7.964	142	43.127
Capital/ha (mil reais)	2.603	5,661	5,584	0,602	56,13
Trab. Tot./ha (horas)	2.603	4,1	14,11	0,723	572,5
De 1000 a menos de 2500 ha					
Nº Estabelecimentos	11.667	9,547	7,924	3	51
VP/ha (mil reais)	1.222	0,564	0,913	0,0232	11,51
Área Prod. (ha)	1.222	11.268	8.688	377	42.819
Capital/ha (mil reais)	1.222	5,874	7,083	0,602	92,94
Trab. Tot./ha (horas)	1.222	3,41	10,11	0,72	220,2
De 2500 ha e mais					
Nº Estabelecimentos	1.337	5,548	3,095	3	20
VP/ha (mil reais)	241	0,871	1,419	0,0257	14,91
Área Prod. (ha)	241	19.943	9.851	54	42.539
Capital/ha (mil reais)	241	5,35	4,239	0,627	24,59
Trab. Tot./ha (horas)	241	6,383	46,05	0,722	713,3

Fonte: Elaborada a partir do Censo Agropecuário de 2006 (IBGE, 2017).

O valor da produção por hectare aumenta da menor faixa de área para a segunda menor. Após isto, decresce conforme o tamanho das fazendas representativas aumentam e aumenta novamente entre as de 1.000 a menos de 2.500 hectares para as de 2.500 hectares e mais. Na média, ele passa de aproximadamente 9 mil reais por hectare (nas fazendas representativas de 0,1 a 0,2 hectares) para 871 reais por hectare (nas de 2.500 e mais hectares). Os valores máximos alcançados são para as fazendas de 0,2 a 3 hectares (em torno de 65 mil reais por hectare).

De maneira semelhante se comportam o capital e o trabalho. O primeiro passa de uma média de 31 mil reais por hectare para 5,3 mil (das fazendas representativas de 0 a 0,1 hectares para de 2.500 e mais). O segundo de 2.378 horas para 6,3 horas por hectare. Estes valores se reduzem na medida que os tamanhos das fazendas representativas aumentam.

De uma maneira geral, os municípios, assim como as fazendas representativas, apresentam heterogeneidade nas variáveis utilizadas na função de produção. Além disto, quando apresentada as estatísticas por fazendas representativas (pelas faixas de área), observa-se que as menores têm maior valor de produção por área, capital por área e trabalho por área. Isto não é suficiente para assumir que a relação é negativa. E por fim, as fazendas representativas de áreas medianas são mais abundantes.

5 RESULTADOS

Inicialmente serão apresentados os resultados relativos aos municípios como observações agregadas, paramétrico e não parametricamente. Regressões para as regiões separadas também foram estimadas. Após estes resultados, as fazendas representativas foram o foco da análise de regressão, no sentido de dar robustez aos resultados.

De uma maneira geral, o modelo teve um bom poder explicativo. A estatística R^2 foi de 0,59, ou seja, o modelo explicou 59% dos valores observados. Este valor é reduzido quando as regiões foram analisadas separadamente. A região Nordeste e a Centro Oeste foram as com menor poder explicativo, com R^2 de 0,44 e 0,32, respectivamente.

A Tabela 4 mostra os resultados da regressão paramétrica para os municípios. A principal variável de interesse, a área produtiva, teve coeficiente 0,027, próximo, porém maior que 0. Isto significa que um aumento em 10% da área vai gerar um aumento de 0,27% do valor de produção por área. De acordo com o atestado por Cornia (1985), Heltberg (1998) e Carletto et al. (2013), como a medida de produtividade é o valor da produção por hectare, isto indica uma rejeição da hipótese da relação inversa entre o tamanho das propriedades agrícolas e a produtividade.

Tabela 4. Regressão paramétrica da produtividade da terra admitindo a função de produção Cobb-Douglas com as regiões representadas por variáveis *dummies*

Variável dependente: VP/ha	Brasil
Área Produtiva	0,027 (2,49)*
Capital/ha	0,803 (48,07)**
Trabalho/ha	0,347 (24,77)**
Norte (<i>dummy</i>)	-0,071 (1,66)
Nordeste (<i>dummy</i>)	0,005 (0,13)
Centro Oeste (<i>dummy</i>)	-0,066 (1,58)
Sul (<i>dummy</i>)	0,154 (5,68)**
Constante	-3,015 (22,34)**
Observações	5193
R^2	0,59

Fonte: Elaborada com os dados do Censo Agropecuário de 2006 (IBGE, 2017).

Nota: Os valores entre parênteses são os desvios padrões dos coeficientes apresentados. *: significante a 10%; **: significante a 5%; ***: significante a 1%.

O capital foi a variável que apresentou maior elasticidade de produção. O seu coeficiente foi de 0,803, ou seja, um aumento de 10% no capital por hectare gera um aumento de 8% no valor de produção por hectare. O mesmo aumento no trabalho, por sua vez, gera 3,47%. Isto é um indicativo de que a produtividade é mais sensível a variações no capital do que ao trabalho, de acordo com os dados observados.

As regiões foram consideradas variáveis *dummies* neste modelo. Greene (2002) explica que estas variáveis comparam alguma característica mantendo-se todo o resto constante. Neste caso, são comparados os valores de produção por área dentre as regiões, mantendo-se todo o resto constante. Elas foram comparadas com a região Sudeste. Então, em condição *coeteris paribus*, as regiões Norte e Centro Oeste têm, respectivamente, -7,1% e -6,6% valor de produção por área, causados por fatores regionais não controlados pelo modelo. O Nordeste apresentou valor próximo a 0 e a Sul (única significativa estatisticamente), 15,6% mais valor de produção por área que o Sudeste.

Estes resultados reforçam a ideia da heterogeneidade das propriedades agrícolas. Além disso, as regiões têm diferentes características físicas causadas pela grande dispersão geográfica. Uma análise de dados em painel controlaria estas diferenças não observadas. Infelizmente, devido à problemas técnicos com a arquitetura do programa de recuperação de dados do Censo Agropecuário do ano de 1975 e 1976 não foi possível a realização do mesmo, a qual seria uma grande contribuição a este trabalho posteriormente. No entanto, as regiões são analisadas isoladamente para verificar se há diferenças na relação entre elas. A Tabela 5 traz estes resultados.

Tabela 5. Regressão paramétrica da produtividade da terra admitindo a função de produção Cobb-Douglas e regiões separadas

Valor da produção/ha	Brasil	Sul	Sudeste	Norte	Nordeste	Centro Oeste
Área	0,027	0,028	0,119	-0,038	-0,025	0,082
Produtiva	(2,50)*	(1,43)	(6,20)**	(1,12)	(1,21)	(2,06)*
Capital/ha	0,840	0,716	1,092	0,303	0,707	0,936
	(76,73)**	(17,10)**	(37,46)**	(5,15)**	(26,35)**	(11,63)**
Trabalho/ha	0,364	0,399	0,215	0,519	0,406	0,153
	(29,84)**	(14,59)**	(8,95)**	(13,71)**	(15,18)**	(2,54)*
Constante	-3,094	-2,812	-4,179	-2,329	-2,609	-3,634
	(23,00)**	(10,73)**	(17,89)**	(5,68)**	(10,14)**	(6,95)**
Observações	5193	1151	1570	394	1677	400
R ²	0,59	0,52	0,51	0,47	0,44	0,32

Fonte: Elaborada com os dados do Censo Agropecuário de 2006 (IBGE, 2017).

Nota: Os valores entre parênteses são os desvios padrões dos coeficientes apresentados. *: significante a 5%; **: significante a 1%.

Os coeficientes relativos às áreas foram significativos apenas para as regiões Sudeste, Centro Oeste e para o Brasil como um todo, o qual não teve alteração. Apesar disso, somente as regiões Norte e Nordeste apresentaram uma relação inversa, ou seja, os coeficientes relativos às suas áreas foram menores que 0.

O capital tem maior elasticidade para as regiões Sudeste e Centro Oeste, enquanto que o trabalho para estas regiões tem as menores. Isto indica que estas regiões apresentam um ganho maior de produtividade advindo do aumento do capital do que as demais. O Norte é a única cuja elasticidade do trabalho é maior que a do capital, portanto apresenta ganhos maiores advindos de um aumento do trabalho. Um aumento em 10% no capital gera um aumento de 3% no valor da produção por hectare, sendo que o mesmo aumento no trabalho gera um aumento de 5,2% no valor da produção por hectare, aproximadamente.

Para as análises não paramétricas os erros padrões foram calculados através do *wild bootstrap* com 500 replicações (HENDERSON; PARMETER, 2015) e apresentados entre parênteses. Eles são aplicados a todas as regressões não paramétricas para obtenção de erros

padrões robustos na presença de heteroscedasticidade (HENDERSON; PARMETER, 2015;). As larguras das janelas foram calculados através do método LSCV, já discutido.

As elasticidades, como fora proposto, são apresentadas por *quantis* (0,25, 0,5, 0,75). A principal variável de interesse apresentou coeficientes significativos no primeiro *quantil* somente para o estimador LQLS, no segundo para o LCLS e LLLS, e no terceiro para o LCLS. Os coeficientes foram próximos entre os diversos estimadores e é um indício de robustez.

A Tabela 6 mostra os resultados da análise de regressão não paramétrica feita para os municípios. Esta não considera a influência da forma funcional para estimar a relação. A primeira observação a ser feita é que o poder explicativo do modelo (R^2) aumentou, ou seja, o modelo não paramétrico teve melhor ajuste que o paramétrico. Vale ressaltar que as observações são os municípios, portanto os *quantis* se referem a eles e não a cada propriedade agrícola isoladamente. A elasticidade de produção da área é negativa no menor *quantil*, se torna positiva e aumenta conforme os *quantis*. Em outras palavras, apenas no primeiro *quantil* a relação foi inversa, e o ganho de produtividade gerado pelo aumento da área é maior nos municípios cuja soma da área das propriedades agrícolas é maior. Portanto, levando em consideração os municípios como observações, uma política de redistribuição de terras poderia ser eficiente no sentido de otimizar a produtividade agropecuária para as propriedades do primeiro *quantil*, cuja relação foi inversa (DEOLALIKAR, 1981; CORNIA, 1985; CHEN et al., 2011).

Tabela 6. Regressão *kernel* não paramétrica utilizando os três estimadores com o valor da produção por hectare como variável dependente

	Área			Cap./ha			Trab./ha			R ²
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	
LCLS	-0,23	0,18	0,675	-1,847	-0,327	2,613	-1,626	-0,317	1,372	0,63
	(-0,348)	(0,091)**	(0,268)**	(0,224)***	(0,051)***	(0,474)***	(0,183)***	(0,14)**	(0,642)**	-
LLS	-0,275	0,287	0,918	-2,141	-0,353	3,227	-1,878	-0,457	1,751	0,62
	(-1,051)	(0,157)*	(-0,837)	(0,203)***	(0,025)***	(0,244)***	(0,166)***	(0,039)***	(0,237)***	-
LQLS	-0,708	0,065	0,865	-2,069	-0,339	2,755	-1,818	-0,452	1,507	0,62
	(0,291)**	(-0,103)	(-1,412)	(0,301)***	(0,023)***	(0,522)***	(0,125)***	(0,071)***	(0,2)***	-

Fonte: Elaborada com os dados do Censo Agropecuário de 2006 (IBGE, 2017).

Nota: Os valores entre parênteses são os desvios padrões dos coeficientes apresentados. *: significativa a 10%; **: significativa a 5%; ***: significativa a 1%.

Q1: *quantil* 0,25; Q2: *quantil* 0,5; Q3: *quantil* 0,7

O coeficiente do capital e do trabalho têm comportamentos semelhantes, ambos apresentam uma relação direta com o aumento do *quantil*. No entanto, a sensibilidade a variações do capital no maior *quantil* é superior que a do trabalho, o que evidencia a importância do capital para estas propriedades, enquanto nos demais elas são equiparadas e inferiores à da área.

A outra forma proposta para visualização dos resultados não paramétricos é a visualização gráfica, mantendo os outros parâmetros fixos em seus valores médios. Os gráficos permitem observar o formato desta relação não observado pela análise das elasticidades.

A Figura 2 demonstra que a relação entre o logaritmo da área (em hectares) e o logaritmo do valor da produção (mil reais) por hectare entre os municípios não é linear. No primeiro gráfico (da esquerda para a direita), relativo ao estimador LCLS, os municípios cujas áreas das propriedades somam entre 403 e 1.100 hectares têm maior valor de produção por hectare.

Figura 2. Relação entre o logaritmo do valor da produção (em mil reais) por hectare pela área (em hectares) para o Brasil para os estimadores LCLS e LLLS

Fonte: Elaborada a partir dos dados do Censo Agropecuário de 2006 (IBGE, 2017).

No segundo, relativo ao LLLS, os municípios que somam 22.000 hectares têm o valor de produção por área máximo. Esta diferença entre os gráficos se dá pela diferença entre os estimadores já comentadas anteriormente. Apesar dela, os formatos de ambos são semelhantes e tem formato de W suavizado, o qual tem uma relação negativa nas menores e alternando entre positiva e negativa.

A não linearidade desta relação é um grande desafio para políticas voltadas a redistribuição de terras no sentido de otimizar a produtividade. Seriam necessárias diferentes políticas para atender as variações da relação conforme os tamanhos das propriedades dos municípios variam.

Para dar robustez aos resultados, as análises serão feitas da mesma maneira, mas com as observações sendo as fazendas representativas. Para investigar se a relação entre os *quantis* é a mesma para todas as regiões, será estimada uma regressão não paramétrica para cada uma delas. Estes resultados são mostrados no capítulo seguinte.

6 ROBUSTEZ DOS RESULTADOS

Até agora os resultados foram apresentados com relação aos municípios como observações. Aqui elas serão as fazendas representativas. Ao utilizá-las, o tamanho da amostra é substancialmente aumentado, e, como já explicitado, isto traz vantagens. Outra vantagem é que a análise econométrica explorou a relação entre as fazendas representativas, e não mais entre os municípios. A apresentação dos resultados foi feita como a anterior. A Tabela 7 traz os resultados relativos ao Brasil e às regiões como variáveis *dummies*. A estatística R^2 sofreu uma pequena redução com relação a Tabela 4. Isso ocorre por causa da mudança das observações.

Tabela 7. Regressão paramétrica utilizando a função de produção Cobb-Douglas e as fazendas representativas como observações

VP/ha	Brasil
Área	0,097 (32,75)**
Capital/ha	0,602 (110,19)**
Trabalho/ha	0,384 (98,83)**
Norte (<i>dummy</i>)	0,003 (0,18)
Nordeste (<i>dummy</i>)	-0,029 (2,33)*
Centro Oeste (<i>dummy</i>)	-0,232 (15,06)**
Sul (<i>dummy</i>)	0,237 (21,82)**
Constante	-3,361 (101,37)**
Observações	58.873
R ²	0,56

Fonte: Elaborada com os dados do Censo Agropecuário de 2006 (IBGE, 2017)

Nota: Os valores entre parênteses são os desvios padrões dos coeficientes apresentados. *: significativa a 10%; **: significativa a 5%; ***: significativa a 1%.

Não há mudanças de sinais nos coeficientes. A elasticidade da área teve um leve aumento, o que reforça o indício da relação direta entre a área e a produtividade. O capital apresentou uma redução de sensibilidade com relação ao valor da produção por hectare. Seu coeficiente passou de 0,803 para 0,602. O trabalho, por sua vez, teve um pequeno aumento.

Com relação às variáveis *dummies*, as fazendas representativas da região Norte tiveram coeficiente próximo de 0. As fazendas nordestinas apresentaram 2,9% de valor de produção a menos que as do Sudeste. As do Centro Oeste foram as mais prejudicadas, o valor de produção por área delas foi 23,2% menor, enquanto que a Sul 23,7% maior.

Ao que tudo indica até o momento, os resultados permanecem sem grandes alterações. Isto indica uma robustez nos resultados, visto que eles são próximos. A Tabela 8 mostra os resultados paramétricos das regiões em regressões separadas.

Tabela 8. Regressão paramétrica utilizando a função de produção Cobb-Douglas para o Brasil e regiões, considerando as fazendas representativas como observações

	Brasil	Sul	Sudeste	Norte	Nordeste	Centro Oeste
Área	0,105 (35,04)**	0,132 (23,41)**	0,146 (25,95)**	0,050 (4,81)**	0,040 (7,68)**	0,070 (6,86)**
Capital/ha	0,649 (164,75)**	0,674 (51,52)**	0,820 (82,45)**	0,309 (16,03)**	0,489 (56,41)**	0,808 (40,00)**
Trabalho/ha	0,384 (105,21)**	0,305 (39,15)**	0,313 (41,02)**	0,581 (42,43)**	0,426 (67,55)**	0,240 (16,38)**
Constante	-3,473 (107,55)**	-3,226 (48,16)**	-3,944 (62,19)**	-3,362 (29,93)**	-3,079 (56,69)**	-3,397 (29,62)**
Observações	58.873	12.557	16.862	4.365	20.511	4.567
R ²	0,55	0,48	0,50	0,63	0,53	0,54

Fonte: Elaborada com os dados do Censo Agropecuário de 2006 (IBGE, 2017).

Nota: Os valores entre parênteses são os desvios padrões dos coeficientes apresentados. *: significante a 10%; **: significante a 5%; ***: significante a 1%.

As regressões para as regiões tiveram melhor ajuste que a Tabela 5. Antes a maior estatística R² era 0,59 e a pior 0,32. Agora, a melhor é 0,63 e a menor 0,48. Além disto, todos os coeficientes foram significativos a 1% de significância. Assim sendo, pode-se afirmar que o modelo que considera estas observações foi mais explicativo para as regiões que ao se considerar as outras.

O coeficiente da variável área da região Sul sofreu um aumento de 0,028 para 0,132 e tornou-se significativo, reafirmando a rejeição da relação inversa. Para o Sudeste e o Centro Oeste a variação foi pequena, enquanto que para o Norte e o Nordeste, o sinal foi alterado e tornaram-se significantes. Contudo, a relação direta continua a ser observada, agora em todas as regiões e, como sugere Deolalikar (1981), Cornia (1985) e Chen et al. (2011) uma

redistribuição de terras, neste caso, não colaboraria para um aumento da produtividade agrícola.

O capital se mostrou bem robusto. A principal diferença foi para o Nordeste, em que seu coeficiente sofreu uma redução de 30% e se aproximou do trabalho. O trabalho, por sua vez, teve um aumento de importância para a região Centro Oeste e continua sendo o maior coeficiente para o Norte, reforçando a importância deste insumo para esta região.

Portanto, até agora, os resultados foram robustos: Houve um aumento da estatística R^2 ; aumento do número e da significância das variáveis; os resultados foram próximos, com apenas algumas alterações; e a relação foi direta para todas as regiões (tornou-se direta e significativa para o Norte e Nordeste).

A próxima bateria de resultados para testar a robustez é com relação à análise não paramétrica. Aqui, além da regressão para o Brasil, como a anterior, faz-se também para cada região separadamente.

A primeira tabela a ser apresentada é a Tabela 9. Apesar da redução do poder explicativo quando comparado ao outro modelo não paramétrico, ele foi bom. Os resultados, de uma maneira geral, também se mostraram robustos. O coeficiente da área, apesar da maioria não ser significativo, apresentou uma tendência de aumento conforme o aumento do *quantil*. A relação se mostrou robusta dentre os *quantis*: Ela foi inversa para o menor, se inverteu para o segundo e foi ainda mais direta para o maior. Assim sendo, a rejeição da evidência da relação inversa feita pelo modelo paramétrico é creditada principalmente pelas propriedades pertencentes aos dois maiores *quantis*. A redistribuição de terras é justificada somente para o menor *quantil*, o qual a relação foi inversa (DEOLALIKAR, 1981; CORNIA, 1985; CHEN et al., 2011).

Tabela 9. Regressão *kernel* não paramétrica utilizando as fazendas representativas como observações e 3 estimadores para o Brasil

	Área			Cap./ha			Trab./ha			R ²
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	
LCLS	-0,245	0,277	1,013	-1,193	0,038	2,583	-0,984	0,005	2,58	0,59
	(-0,677)	(1,662)*	(0,205)***	(0,143)***	(0,01)***	(0,495)***	(0,257)***	(-0,011)	(0,098)***	-
LLLS	-0,406	0,289	1,236	-1,275	0,064	2,702	-1,021	0,094	2,994	0,59
	(-0,339)	(-0,199)	(0,305)***	(0,201)***	(-0,054)	(0,289)***	(0,241)***	(-0,395)	(0,855)***	-
LQLS	-0,551	0,258	1,201	-1,282	0,022	2,535	-1,068	0,018	2,869	0,6
	(-0,609)	(-0,365)	(-0,852)	(-0,872)	(0,006)***	(0,659)***	(0,415)**	(0,006)***	(0,2070)***	-

Fonte: Elaborada com os dados do Censo Agropecuário de 2006 (IBGE, 2017).

Nota: Os valores entre parênteses são os desvios padrões dos coeficientes apresentados. *: significante a 10%; **: significante a 5%; ***: significante a 1%.

Q1: *quantil* 0,25; Q2: *quantil* 0,5; Q3: *quantil* 0,75

O capital também apresentou resultados robustos. A maior diferença de coeficiente foi para o Q2, em que o sinal se inverteu. O aumento da sensibilidade com relação ao valor da produção por área conforme o aumento dos *quantis* continua a ser observada. Além disto, para o Q3, o capital apresentou uma elasticidade muito maior que com relação a área, ou seja, um aumento no capital destas propriedades gera mais que o dobro de aumento de produtividade do que na área. Para os demais *quantis*, o contrário acontece, a área é o insumo cuja variação tem maior impacto positivo na produtividade, da mesma maneira que na Tabela 6. O trabalho apresentou resultados semelhantes ao capital, tanto em magnitude, como em tendência. Apesar da redistribuição de terras contribuir apenas para o aumento da produtividade agrícola entre as propriedades do primeiro *quantil*, um aumento na área continua a ser mais benéfico que o aumento dos demais insumos para os dois últimos *quantis*.

As próximas tabelas são referentes à análise não paramétrica para cada região. O número de observações de cada uma são iguais aos da Tabela 8. A região Sudeste é representada pela Tabela 10. Seus coeficientes, em suma, são semelhantes aos do Brasil como um todo. A principal variável de interesse tem elasticidade negativa no Q1 para todos os estimadores, não rejeitando a relação inversa (DEOLALIKAR, 1981; HELTBERG, 1998; CARLETTO et al., 2013). No Q2 ela passa a ser positiva, ou seja, a relação se torna direta e é ainda maior no Q3, como o ocorrido na tabela anterior.

Tabela 10. Regressão *kernel* não paramétrica considerando as fazendas representativas como observações para a região Sudeste

	Área			Cap./Área			Trab./Área			R ²
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	
LCLS	-0,544	0,344	1,406	-2,345	1,306	3,883	-0,666	0,815	2,513	0,55
	(0,311)*	(-1,225)	(0,246)***	(0,321)***	(0,45)***	(0,73)***	(0,06)***	(0,474)*	(0,3)***	-
LLLS	-0,811	0,419	1,724	-2,488	1,436	3,878	-0,655	0,94	2,893	0,55
	(0,469)*	(0,163)**	(0,171)***	(0,088)***	(0,259)***	(0,387)***	(0,039)***	(0,06)***	(0,458)***	-
LQLS	-0,931	0,399	1,649	-2,525	1,308	3,937	-0,711	0,655	2,652	0,56
	(0,501)*	(-0,285)	(0,136)***	(0,414)***	(0,244)***	(0,433)***	(0,086)***	(-0,886)	(0,383)***	-

Fonte: Elaborada com os dados do Censo Agropecuário de 2006 (IBGE, 2017).

Nota: Os valores entre parênteses são os desvios padrões dos coeficientes apresentados. *: significativa a 10%; **: significativa a 5%; ***: significativa a 1%.

Q1: *quantil* 0,25; Q2: *quantil* 0,5; Q3: *quantil* 0,75

A tendência de aumento de elasticidades conforme o aumento do *quantil* persiste. O capital, apesar do comportamento ser parecido com o do Brasil dentre os *quantis*, tem uma importância significativamente maior para o percentil 50, quando comparado a Tabela 9. Antes os coeficientes eram próximos de 0, agora eles estão entre 1,3 e 1,44, o que significa que uma variação em sua quantidade tem maior impacto no valor da produção por hectare nessas propriedades. Assim como o capital, o trabalho teve o mesmo comportamento com uma diferença para o primeiro *quantil*, o qual a aproximação dos coeficientes da área e do trabalho apontam para uma convergência de importância destes insumos para o valor da produção por hectare.

Portanto, o aumento da quantidade de insumos para as maiores propriedades da região Sudeste é mais produtivo do que para as menores. Isto pode ser observado devido à grande discrepância das elasticidades das variáveis do maior *quantil* em relação aos dois menores. As medianas desta região, pelo mesmo motivo, têm um aumento de produtividade maior gerado pelo aumento dos insumos que as brasileiras. Além disto, além da área, o trabalho também se mostrou o insumo de maior sensibilidade de variação com relação a produtividade para as propriedades do último *quantil*.

A Tabela 11 mostra que as elasticidades da variável área para o Sul diferem das da região Sudeste. Apesar de não haver significância para nenhum estimador do menor *quantil*, ele apresenta uma relação direta próxima a 0 que se mantém para os maiores. Segundo estes resultados, a redistribuição de terras não seria eficiente para nenhum *quantil* desta região, porque as elasticidades de produção da área são positivas para todos os *quantis*, evidência que rejeita a relação inversa (DEOLALIKAR, 1981; HELTBERG, 1998; CARLETTO et al., 2013) e, portanto, a redistribuição de terras poderia não gerar ganhos de produtividade (DELOLAIKAR, 1981; CORNIA, 1985, CHEN et al., 2011).

Tabela 11. Regressão *kernel* não paramétrica considerando as fazendas representativas como observações para a região Sul

	Área			Cap./Área			Trab./Área			R ²
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	
LCLS	0,037	0,381	0,732	0,491	2,152	3,235	-0,115	0,97	1,76	0,52
	(-0,493)	(-1,285)	(-2,66)	(-0,819)	(0,905)**	(0,34)***	(0,07)*	(0,502)*	(0,58)***	-
LLLS	0,043	0,628	1,207	0,624	2,351	3,912	-0,083	1,257	2,262	0,53
	(-0,09)	(0,339)*	(0,282)***	(-1,153)	(0,189)***	(0,247)***	(0,049)*	(0,298)***	(0,517)***	-
LQLS	-0,09	0,557	1,233	0,229	2,175	3,708	-0,223	1,281	2,373	0,54
	(-0,166)	(0,111)***	(0,079)***	(-0,7)	(0,269)***	(0,871)***	(0,074)***	(0,192)***	(1,082)**	-

Fonte: Elaborada com os dados do Censo Agropecuário de 2006 (IBGE, 2017).

Nota: Os valores entre parênteses são os desvios padrões dos coeficientes apresentados. *: significativa a 10%; **: significativa a 5%; ***: significativa a 1%.

Q1: *quantil* 0,25; Q2: *quantil* 0,5; Q3: *quantil* 0,75

Nota-se um aumento das elasticidades do capital e do trabalho para os dois menores *quantis* quando comparado com o Sudeste e com o Brasil. O capital é a variável que apresenta maior diferença, mas também segue o padrão de aumento de coeficiente conforme o aumento do *quantil*. Apesar de não significativo, ele deixa de ser negativo para o primeiro *quantil*. No percentil 50 sua elasticidade aumenta consideravelmente, ou seja, uma variação dele nessas propriedades gera mais valor de produção por hectare que as propriedades do mesmo *quantil* do Sudeste. Outra diferença relevante é que o capital, seguido do trabalho, são os insumos cuja variação tem maior impacto na produtividade para o segundo *quantil* e o capital, seguido da área, para o primeiro.

A próxima região a ser analisada é a Centro Oeste (Tabela 12). O coeficiente de todas as variáveis mantém o mesmo padrão de comportamento entre os *quantis*. A relação entre a área e a produtividade é inversa para os dois menores e direta para o maior, que tem a menor elasticidade entre todas as regiões analisadas até agora. A redistribuição de terras deveria ser considerada, pelo mesmo motivo já explicado nas outras regiões, para os dois menores *quantis*.

Tabela 12. Regressão *kernel* não paramétrica considerando as fazendas representativas como observações para a região Centro Oeste

	Área			Cap./Área			Trab./Área			R ²
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	
LCLS	-0,618	-0,189	0,154	-2,274	-1,313	-0,229	-0,937	-0,087	0,033	0,6
	(-0,892)	(-0,172)	(-1,74)	(0,155)***	(0,302)***	(-1,112)	(0,18)***	(-0,054)	(0,018)*	-
LLLS	-1,27	-0,511	-0,014	-2,662	-1,377	-0,328	-1,238	-0,134	0,008	0,6
	(0,358)***	(0,145)***	(-0,148)	(0,161)***	(0,084)***	(0,052)***	(0,145)***	(0,082)*	(0,003)***	-
LQLS	-1,638	-0,562	0,102	-2,511	-1,28	-0,048	-1,374	-0,201	0,016	0,62
	(0,708)**	(0,342)*	(-0,168)	(0,503)***	(0,101)***	(-0,07)	(-1,125)	(0,055)***	(-0,037)	-

Fonte: Elaborada com os dados do Censo Agropecuário de 2006 (IBGE, 2017).

Nota: Os valores entre parênteses são os desvios padrões dos coeficientes apresentados. *: significativa a 10%; **: significativa a 5%; ***: significativa a 1%.

Q1: *quantil* 0,25; Q2: *quantil* 0,5; Q3: *quantil* 0,75

O capital se mostra negativo para todos os estimadores e *quantis*, enquanto o trabalho é positivo, porém próximo a 0 no último *quantil*. O valor da produção por hectare é mais sensível positivamente a variações no trabalho e na área, variando dentre os estimadores, do que no capital nos dois primeiros *quantis*, enquanto o terceiro é mais sensível ao trabalho. Isto indica que essa região é mais dependente da área e do trabalho do que o capital.

Para a região Norte (Tabela 13) o coeficiente da área é negativo para os dois primeiros *quantis* e, como já explicitado, isto indica uma relação inversa, que se torna direta para o último. O trabalho se mostrou o insumo mais importante para o último *quantil*, porque sua elasticidade de produção é a maior, seguida do capital e da área, respectivamente, ou seja, um aumento no trabalho gera mais valor da produção por hectare que o capital e o trabalho. Para o segundo, a área foi mais importante, apesar de não significativa, pois ela foi o insumo com maior elasticidade; para o primeiro o capital e a área têm a mesma importância. A redistribuição de terras, pelo mesmo motivo do Centro Oeste, poderia gerar ganhos de produtividade para os dois primeiros *quantis*. Além disto, essa região tem elasticidade do capital com relação a produção superior que o Centro Oeste em todos os *quantis*, porém, muito inferior ao Brasil e as demais regiões analisadas.

Tabela 13. Regressão *kernel* não paramétrica considerando as fazendas representativas como observações para a região Norte

	Área			Cap./Área			Trab./Área			R ²
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	
LCLS	-0,34	-0,039	0,34	-0,334	-0,106	0,2	-1,433	-0,118	3,24	0,68
	(-0,262)	(-0,116)	(0,143)**	(0,091)***	(0,015)***	(-0,189)	(0,143)***	(0,034)***	(1,808)*	-
LLLS	-0,473	-0,034	0,417	-0,619	-0,179	0,276	-1,662	-0,149	3,757	0,68
	(0,245)**	(-0,304)	(-0,388)	(0,102)***	(0,038)***	(-0,429)	(0,133)***	(0,019)***	(0,637)***	-
LQLS	-0,601	-0,048	0,528	-0,592	-0,143	0,27	-1,546	-0,165	3,297	0,7
	(0,318)*	(-0,125)	(0,106)***	(0,044)***	(0,049)***	(-3,875)	(0,203)***	(0,018)***	(1,277)***	-

Fonte: Elaborada com os dados do Censo Agropecuário de 2006 (IBGE, 2017).

Nota: Os valores entre parênteses são os desvios padrões dos coeficientes apresentados. *: significativa a 10%; **: significativa a 5%; ***: significativa a 1%.

Q1: *quantil* 0,25; Q2: *quantil* 0,5; Q3: *quantil* 0,75

A última região é a Nordeste, e seus resultados estão na Tabela 14. Como observado até agora, as variáveis aumentaram o coeficiente com o aumento do *quantil*. A relação entre a área e a produtividade é inversa somente para o primeiro *quantil*. A área é o insumo de maior elasticidade para os dois primeiros *quantis*, enquanto o trabalho é para o maior, ou seja, um aumento na área é mais benéfico que o aumento nos demais insumos para os dois menores *quantis* e o trabalho é mais benéfico para o maior.

Tabela 14. Regressão *kernel* não paramétrica considerando as fazendas representativas como observações para a região Nordeste

	Área			Cap./Área			Trab./Área			R ²
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	
LCLS	-0,068	0,237	0,798	-0,73	-0,104	0,936	-1,832	-0,474	2,765	0,57
	(-0,141)	-0,306	(0,112)***	(-0,579)	(0,003)***	(-0,768)	(0,489)***	(0,047)***	(0,800)***	-
LLLS	-0,236	0,241	0,992	-0,739	-0,095	0,981	-1,76	-0,459	3,554	0,57
	(-0,207)	(-0,161)	(-0,812)	(0,044)***	(0,019)***	(-0,798)	(0,232)***	(0,071)***	(0,647)***	-
LQLS	-0,708	0,065	0,865	-2,069	-0,339	2,755	-1,818	-0,452	1,507	0,62
	(0,291)**	(-0,103)	(-1,412)	(0,301)***	(0,023)***	(0,522)***	(0,125)***	(0,071)***	(0,2)***	-

Fonte: Elaborada com os dados do Censo Agropecuário de 2006 (IBGE, 2017).

Nota: Os valores entre parênteses são os desvios padrões dos coeficientes apresentados. *: significativo a 10%; **: significativo a 5%; ***: significativo a 1%.

Q1: *quantil* 0,25; Q2: *quantil* 0,5; Q3: *quantil* 0,75

As regiões são heterogêneas entre si, fato ainda mais estilizado pelas diferenças de elasticidades dentre elas para todas as variáveis e *quantis* dentre as regiões (e heterogêneas entre os *quantis*). Por isto, no que tange a políticas, tanto de redistribuição de terras, quanto a outras voltadas a agropecuária, deveria levar em consideração esta heterogeneidade e diversificá-las de acordo com as necessidades de cada região e tamanho das propriedades, o que é um desafio pela grande quantidade de fazendas e extensão territorial brasileiro. Apesar disto, observa-se um padrão entre variáveis: A elasticidade de produção delas é diretamente proporcional aos *quantis*.

Se a elasticidade produção de todos os insumos das maiores propriedades são superiores aos das menores, então um aumento deles geraria maior valor de produção que nas menores propriedades. Em outras palavras, a produtividade das propriedades do maior *quantil* são mais sensíveis positivamente a alterações nos insumos que as dos menores.

O Brasil e as regiões Sudeste e Nordeste têm a relação entre a área e a produtividade inversa no primeiro *quantil*, enquanto a Centro Oeste e a Norte inversa nos dois primeiros *quantis* e a Sul direta em todos. A magnitude da relação direta das propriedades (cuja elasticidade foi positiva) foi o suficiente para fazer com que a relação fosse direta na análise paramétrica (global), portanto pode-se afirmar que as maiores propriedades são as responsáveis por essa relação direta.

O último exercício de robustez é a análise contra fatorial entre as áreas produtivas e o valor da produção por hectare, ambos em logaritmos. Os gráficos também são apresentados por região. Eles foram agrupados por estimadores¹³.

As Figuras a seguir (3 a 5) mostram os gráficos para o Brasil e regiões, com relação ao estimador LCLS. Em suma, como já observado anteriormente nas regressões, o formato da relação entre o logaritmo natural da área e o logaritmo natural do valor da produção por hectare difere dentre as regiões, não é monotônica e nem linear. Isto explicita a heterogeneidade das fazendas representativas dentre as regiões. Para transformar os valores dos logaritmos naturais é necessário elevar o número de Euler (explicitado pela letra *e*) pelo valor do logaritmo que se deseja transformar.

A Figura 3 apresenta os gráficos para o Brasil e para o Sul. O formato da relação para o Brasil, mantendo as outras variáveis em seus valores médios, se assemelha ao gráfico relativo aos municípios. As menores propriedades, e as maiores, tem maior valor da produção

¹³ O estimador LQLS foi abtido por causa do alto custo computacional explicado na metodologia.

por hectare. Percebe-se também que a produtividade marginal da terra é decrescente para as menores propriedades e se torna crescente para as médias e grandes (passa por um pequeno decréscimo nas maiores).

Figura 3. Relação entre o logaritmo natural do valor da produção (em mil reais) por hectare e a área (em hectares) para o Brasil e o Sul mantendo-se fixas as outras variáveis (LCLS)

Fonte: Elaborada com os dados do Censo Agropecuário de 2006 (IBGE, 2017).

Há uma grande diferença entre os formatos das regiões para com o Brasil. As maiores propriedades da região Sul têm maior valor de produção por área que as menores. Observa-se que o valor da produção por hectare das fazendas compreendidas entre o logaritmo natural de 2 a 8 (7,4 a 2980 hectares), com o trabalho e o capital fixos em seus valores médios, aumenta conforme há um aumento da área. O contrário ocorre para as demais propriedades.

O Centro Oeste é semelhante ao Sul, as maiores fazendas representativas apresentaram vantagens produtivas em relação as menores. Mas a diferença é menos abrupta na região Centro Oeste. Porém as menores fazendas têm uma significativa desvantagem. Além disso, as fazendas de até 55 hectares têm um ganho de valor de produção por aumento de área maior que as demais, mantendo-se o capital e o trabalho constantes em seus valores médios.

Figura 4. Relação entre o valor da produção (em mil reais) por hectare pela área (hectares) para o Centro Oeste e o Nordeste mantendo-se as outras variáveis fixas (LCLS)

Fonte: Elaborada com os dados do Censo Agropecuário de 2006 (IBGE, 2017).

O Nordeste tem um formato de U invertido. As menores e as maiores fazendas representativas têm menor valor de produção que as medianas. Apenas as propriedades compreendidas entre 7,4 e 54 hectares aumentaram seu valor de produção com o aumento da área.

A Figura 5 representa a mesma relação para as regiões Norte e Sudeste. As propriedades nortistas têm maior valor de produção por unidade de área que as demais. No entanto, estas apresentaram uma relação fortemente negativa entre a área e o valor de produção por hectare (com os valores do capital e trabalho fixos em seus valores médios). As propriedades acima de 54 hectares (logaritmo natural de 4) sofrem alterações menos representativas com o aumento da área sob as mesmas circunstâncias.

Figura 5. Relação entre o valor da produção (em mil reais) por hectare pela área (em hectares) para o Norte e o Sudeste mantendo-se as outras variáveis fixas (LCLS)

Fonte: Elaborada com os dados do Censo Agropecuário de 2006 (IBGE, 2017).

O Sudeste, por fim, tem um formato semelhante ao Sul. Quando o capital e o trabalho são considerados fixos em seus valores médios, a relação predominante entre a área e o valor de produção por hectare é positiva, salvo para as fazendas compreendidas entre 3 e 20 hectares e entre 3000 e 9000 hectares, em que ela é negativa. Além disto, observa-se uma superioridade de valor de produção por área nas maiores propriedades.

A Figura 4 apresenta os mesmos gráficos para o estimador LLS. De uma maneira geral, os formatos não sofreram grandes alterações. Henderson e Parmeter (2014) chamam a atenção para a melhor captação de não linearidades do estimador LLS em relação ao LCLS devido as aproximações de Taylor feitas entre os pontos, ao contrário da constante aplicada pelo LCLS.

Figura 6. Relação entre o valor da produção (em mil reais) por hectare pela área (em hectares) para o Brasil e regiões mantendo-se as outras variáveis fixas (LLS)
 Fonte: Elaborada com os dados do Censo Agropecuário de 2006 (IBGE, 2017).

A maior diferença observada foi para o Cento Oeste. A superioridade de valor de produção por hectare das maiores fazendas representativas foi maior neste estimador. Porém, os formatos de todas as regiões não se alteraram substancialmente.

Os resultados apresentados nesta seção condizem com os apresentados na seção anterior. Isto significa que os resultados foram robustos. A relação inversa foi rejeitada na regressão paramétrica para todas as regiões e para o Brasil.

A não paramétrica apresentou a elasticidade por *quantis*. Os resultados também foram convergentes. A relação inversa e, portanto, o potencial de ganho de produtividade gerado pela redistribuição de terras (DEOLALIKAR, 1981; CORNIA, 1985; CHEN et al., 2011) foi observado entre as propriedades do menor *quantil* de todas as regiões e do segundo *quantil* das regiões Norte e Centro Oeste. Além disto, foi observada uma heterogeneidade, que já era esperada, entre as regiões. Cada região apresentou diferentes elasticidades referentes aos insumos e aos *quantis* e diferentes magnitudes da relação inversa.

A análise gráfica permitiu visualizar que a relação é direta para as maiores e inversa para as menores propriedades, as quais apresentou maior valor da produção por hectare para o Brasil, o que variou entre as regiões: A relação foi inversa para as maiores propriedades da região Sul e para as pequenas e grandes do Sudeste, mas as maiores propriedades tiveram maior valor de produção para ambas; a relação foi direta para as menores, tornando-se aproximadamente constante para as demais no Centro Oeste que, além das maiores, as medianas também apresentaram valor da produção superior, as quais foram as mais vantajosas para o Nordeste, cujo formato da relação se assemelha ao U invertido; e para o Norte as menores foram mais produtivas e o formato da relação é próximo a um L, sendo negativo para as menores e constante para as demais.

Em suma, a relação inversa não é linear e nem monotônica. Ela foi inversa para o primeiro *quantil* de todas regiões e para o segundo das regiões Norte e Centro Oeste. No entanto, somente a relação inversa destes *quantis* não foi o suficiente para fazer com que a estimativa global paramétrica também fosse inversa.

Não é possível encontrar justificativas para estes fatos através dos dados e do modelo utilizado. Vale ressaltar que a análise não paramétrica passa pelo problema da dimensionalidade, mas através dele foi possível entender como se dá a relação de forma a evitar qualquer viés causado pela forma funcional. Em outras palavras, a verdadeira relação entre o tamanho das propriedades agrícolas e o valor da produção por hectare foi estimada somente de acordo com os próprios dados.

7 CONCLUSÕES

Verificou-se comum a utilização da análise de regressão para estimar funções de produção com o intuito de analisar a relação entre o tamanho das propriedades agrícolas e a produtividade. O emprego de uma função de produção requer a decisão de uma forma funcional, que pode causar viés nesta relação (ULVELING; FLETCHER, 1970). Alguns autores propõe o uso da análise não paramétrica (VERSCHELDE et al., 2013) para esta finalidade. Motivado por Henderson e Parmeter (2014), que propõe o uso da análise de regressão *kernel* não paramétrica para evitar vieses das formas funcionais e aconselham o uso dela para a estimação de funções de produção, propôs-se estimar esta relação para o Brasil através dos dados do Censo Agropecuário de 2006 fornecido pelo IBGE. A estratégia empregada consistiu em fazer a análise paramétrica, cujo estimador “enxerga” os dados de uma forma global e a não paramétrica, que analisa os dados localmente e que ainda não foi utilizada para este fim para o Brasil. Utilizou-se duas agregações diferentes fornecidas pelo IBGE: Por município agregado e por faixa da área por município (fazendas representativas). As regiões separadas também foram analisadas.

A análise paramétrica mostrou que a relação é direta e próxima a 0 para o Brasil e regiões em ambas as agregações porque os coeficientes estimados para a área são significativamente maiores que 0 (DEOLALIKAR, 1981; HELTBERG, 1998; CARLETTO et al., 2013). A não paramétrica explicitou as diferenças de elasticidades entre os *quantis*, cujo padrão de aumento das elasticidades de todas as variáveis conforme o aumento dos *quantis* indicou que a produtividade das maiores propriedades é mais sensível às alterações nos insumos, ou seja, elas têm maiores ganhos advindos do aumento da quantidade de insumos que as menores. A relação foi inversa para o menor *quantil* do Brasil, porque o coeficiente da área estimado foi negativo (DEOLALIKAR, 1981; HELTBERG, 1998; CARLETTO et al., 2013), ou seja, a hipótese da relação inversa entre a área e a produtividade é confirmada para as menores propriedades, porém rejeitada para as médias e grandes. Dentre as regiões a relação variou: Ela foi inversa para o menor *quantil* de todas as regiões, para os dois primeiros do Norte e Centro Oeste e direta para todos do Sul. Além disso, a variação no valor de produção tem variações diferentes entre os *quantis* de cada região. A relação direta pode ser creditada às maiores propriedades, fato estilizado pela análise não paramétrica.

As análises gráficas, que mantiveram as outras variáveis constantes em seus valores médios, permitiram observar que a relação em questão não é linear, não é monotônica e difere

entre as regiões, sendo um desafio para a criação de políticas relacionadas com a redistribuição de terras. Para o Brasil, o formato da relação é semelhante a um W. Para o Sul ela somente foi negativa para as maiores propriedades. O Sudeste, cuja relação foi predominantemente positiva, apresentou trechos negativos para as pequenas e grandes propriedades. No Nordeste ela teve formato de U invertido, enquanto que no Centro Oeste ela foi positiva para as menores propriedades, aproximadamente constante para as médias e grandes tornando-se negativa para as maiores. O Norte, por fim, teve uma relação em formato de L, sendo extremamente negativa para as menores.

Com relação às propostas de políticas relacionadas, do ponto de visto da produtividade, a redistribuição de terras seria justificada entre as propriedades dos *quantis* cuja relação foi inversa, porque, como afirmado por Deolalikar (1981), Cornia (1985) e Chen et al. (2011), a relação inversa teria potencial de gerar ganhos de produtividade advindos da redistribuição de terras. Além do mais, não é proposto nenhum outro tipo de política, pois como Henderson e Parmeter (2014) enfatizam, o uso da análise não paramétrica não é o mais indicado para propor políticas públicas, devido à dificuldade de interpretação dos resultados.

Para trabalhos futuros, o uso de um nível de desagregação por fazenda (ao nível de microdados) seria benéfico para a análise não paramétrica e a utilização de dados em painel com dados do Censo Agropecuário de 1995 e 1996 e de um possível futuro Censo Agropecuário permite o controle de características não observadas o que traz vantagens para a análise.

REFERÊNCIAS

- ALMEIDA, A.N. **Is there a relationship between farm Size and farm income in Nicaragua?** A panel data analysis with non-parametric methods. 2016. 38 p.
- ASSUNÇÃO, J.J.; BRAIDO, L.H.B. Testing household-specific explanations for the inverse productivity relationship. **American Journal of Agricultural Economics**, Oxford, v. 89, n. 4, p. 980-990, nov. 2007.
- ASSUNÇÃO, J.J.; GHATAK, M. Can unobserved heterogeneity in farmer ability explain the inverse relationship between farm size and productivity. **Economics Letters**, Amsterdam, v. 80, n. 3, p. 189-94, ago. 2003.
- ALI, A.D.; DEININGER, K. Is there a farm size-productivity relationship in African agriculture? Evidence from Rwanda. **Land Economics**, Madison, v. 91, n. 2, p. 317-343, mai 2015.
- AITCHISON, J.; AITKEN, C.G.G. Multivariate binary discrimination by kernel method; **Biometrika**, Oxford, v. 63, n. 3, p. 413-420, dez. 1976.
- BARRET, C.B. On price risk and the inverse farm size-productivity relationship. **Journal of Development Economics**, Amsterdam, v. 51, n. 2, p. 193-215, dez. 1996.
- BARRETT, C.B.; BELLEMARE, M.F.; HOU, J.Y. Reconsidering conventional explanations of the inverse productivity-size relationship. **World Development**, Amsterdam, v. 38, n. 1, p. 88-97, jan. 2010.
- BENJAMIN, D. Can unobserved land quality explain the inverse productivity relationship? **Journal of Development Economics**, Amsterdam, v. 46, n. 1, p. 51-84, fev. 1995.
- BERAN, R. Minimum distance procedure. In: **Handbook of statistics**. Amsterdam: North-Holland Publishing Co., 1984. cap. 30, p. 741-751, v. 4.
- BERRY, R.A.; CLINE W.R. **Agrarian structure and productivity and developing countries**. Baltimore: Johns Hopkins University Press, 1979, 260 p.
- BHALLA, S. Does quality matter?: Theory and measurement. **Journal of Development Economics**, v. 29, n. 1, p. 45-62, jul. 1988.
- BINSWANGER, H.P.; DEININGER, K.; FEDER, G. Power, distortions, revolt and reforming agricultural land relations. In: **Handbook of Development Economics**. 9. ed. Amsterdam: North-Holland Publishing Co., 1995. cap. 42, p. 2659-2772, v. 3B.
- BINSWANGER, H.P.; ROSENZWEIG M.R. Behavioural and material determinants of production relations in agriculture. **Journal of Development Studies**, Londres, v. 22, n. 3, p. 503-539, jan. 1986.
- CARLETTO, C.; SAVASTANO, S.; ZEZZA, A. Fact or artifact: the impact of measurement errors on the farm size-productivity relationship. **Journal of Development Economics**, Amsterdam, v. 103, n. C, p. 254-261, jul. 2013.

CARTER, M.R.; WIEBE, K. Access to capital and its impact on agrarian structure and productivity in Kenya. **American Journal of Agricultural Economics**, Milwaukee, v. 72, n. 5, p. 1146-1150, dez. 1990.

CHAYANOV, A.V. The theory of peasant economy, 1926. Editado por: D. Thorner, B. Kerblay, e R. E. F. Smith. Homewood: **Irwin Publishers**, 1966. 317 p.

CHEN, Z.; HUFFMAN, W.E.; ROZELLE, S. Inverse relationship between productivity and farm size: the case of China. **Contemporary Economic Policy**, Morgantown, v. 29, n. 4, p. 580-592, out. 2011.

CORNIA, G.A. Farm size, land yields and the agricultural production function: an analysis for fifteen developing countries. **World Development**, Amsterdam, v. 13, n. 4, p. 513-534, abr. 1985.

DEATON, A. Rice prices and income distribution in Thailand: a non-parametric analysis. **The Economic Journal**, Nova Jersey, v. 99, n. 395, p. 1-37, 1989.

DEININGER, K.; D. BYERLEE. The rise of large farms in land abundant countries: do they have a future? **World Development**, Amsterdam, v. 40, n. 4, p. 701-714, abr. 2012.

DEININGER, K.; CARLETTO, G.; SAVASTANO, S. Land fragmentation, corpland abandonment, and land market operation in Albania. **World Development**, Amsterdam, v. 40, n. 10, p. 2108-2122, out. 2012.

DEOLALIKAR, A.B. The inverse relationship between productivity and farm size: a test using regional data from India. **American Journal of Agricultural Economics**, Oxford, v. 63, n. 2, p. 275-279, mai. 1981.

DORWARD, A. Farm size and productivity in malawian smallholder agriculture. **Journal of Development Studies**, Londres, v. 35, n. 5, p. 141-61, abr. 1999.

EUBANK, R.L. **Nonparametric regression and spline smoothing**. 2. ed. Nova Iorque: Marcel Dekker, 1999. 355 p. v. 157.

FEDER, G. The relation between farm size and farm productivity: the role of family labor, supervision and credit constraints. **Journal of Development Economics**, Amsterdam, v. 18, n. 2-3, p. 297-313, ago. 1985.

FINAN, F.; SADOULET, E.; DE JANVRY, A. Measuring the poverty reduction potential of land in rural Mexico. **Journal of Development Economics**, Amsterdam, v. 77, n. 1, p. 27-51, jun. 2005.

GASQUES, J.G.; BASTOS, E.T.; BACCHI, M.R.P.; VALDES, C. Produtividade total dos fatores e transformações da agricultura brasileira: análise dos dados dos censos agropécuarios. In: SOCIEDADE BRASILEIRA DE ECONOMIA, ADMINISTRAÇÃO E SOCIOLOGIA RURAL, Campo Grande. **Anais do XXXVIII Encontro Nacional de Economia**. P. 1-21, 2009.

GREENE, W.H. **Econometric analysis**. 5. ed. Nova Jersey: Pearson Education, 2002. 802 p.

HELFAND, S.M.; LEVINE, E.S. Farm size and the determinants of productive efficiency in the Brazilian Center-West. **Agricultural Economics**, Nova Jersey, v. 31, n. 2-3, p. 241-249, dez. 2004.

HELFAND, S.M.; MAGALHÃES, M.M.; RADA, N.E. **Brazil's agricultural total factor productivity growth by farm size**. Washington D. C.: IDB, 2015, n.609, 63 p. (Working Paper Series, 609).

HELTBERG, R. Rural market imperfections and the farm size-productivity relationship: evidence from Pakistan. **World Development**, Amsterdam, v. 26, n. 10, p. 1807-1826, out. 1998.

HENDERSON, H. Considering technical and allocative efficiency in the inverse farm size-productivity relationship. **Journal of Agricultural Economics**, Nova Jersey, v. 66, n. 2, p. 442-469, jun. 2015.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. **Censo Agropecuário 2006**: Brasil, Grandes Regiões e Unidades da Federação. 2. ed. Rio de Janeiro: IBGE, 2011. 744 p.

_____. **Sistema IBGE de Recuperação Automática – SIDRA**, 2016. Disponível em: <<https://sidra.ibge.gov.br/pesquisa/censo-agropecuaria/censo-agropecuario-2006/segunda-apuracao>> Acesso em: 11 out. 2016.

JACOBY, H.G. Shadow wages and peasant family labour supply: an econometric application to the Peruvian Sierra. **Review of Economic Studies**, Oxford, v. 60, n. 4, p. 903–921, out. 1993.

KAWASAKI, K. The costs and benefits of land fragmentation of rice farms in Japan. **Australian Journal of Agricultural and Resource Economics**, Nova Jersey, v. 54, n. 4, p. 509-526, out. 2010.

KIMHI, A. Plot size and maize productivity in Zambia: is there an inverse relationship? **Agricultural Economics**, Nova Jersey, v. 35, n. 1, p. 1–9, jul. 2006.

LAMB, R.L. Inverse productivity: land quality, labor markets, and measurement error. **Journal of Development Economics**, Amsterdam, v. 71, n. 1, p. 71-95, jun. 2003.

LI, Q.; RACINE, J.; WOOLDRIDGE. Efficient estimation of average treatment effects with mixed categorical and continuous data. **Journal of Business & Economic Statistics**, Londres, v. 27, n. 2, p. 206-223, 2009.

MENDES M.S.; TEXEIRA C.E.; SALVATO M.A. Investimentos em infra-estrutura e produtividade total dos fatores na agricultura brasileira: 1985–2004. **Revista Brasileira de Economia**, Rio de Janeiro, v. 63, n. 2, p. 91-102, abr./jun. 2009.

RADA, N.E.; BUCCOLA, S.T. Agricultural policy and productivity: evidence from Brazilian censuses. **Agricultural Economics**, Nova Jersey, v. 43, n. 4, p. 355–367, jul. 2012.

RADA, N.; VALDES, C. Policy, technology and efficiency of Brazilian agriculture. **Economic Research Service**, Washington, n. 137, p. 1-30, 2012.

SEN, A.K. An aspect of Indian agriculture. **The Economic Weekly**, Mumbai, p. 243-246, fev. 1962.

SEN, A.K. **Employment, technology, and development**. Oxford: Clarendon Press, 1975. 193 p.

ULVELING, F.E.; FLETCHER, L.B. A Cobb-Douglas production function with variable returns to scale. **American Journal of Agricultural Economics**, Oxford, v. 52, n. 2, p. 322-326, mai. 1970.

VICENTE, J.R.; ANEFALOS, L.C.; CASER D.V. Produtividade agrícola no Brasil, 1970-1995. In: Congresso Brasileiro de Economia e Sociologia Rural, 38., 2000, Rio de Janeiro. **Produtividade agrícola no Brasil, 1970-1995**, São Paulo, 2001. p. 33-55. Disponível em: <https://www.researchgate.net/profile/Jose_Vicente10/publication/262524562_Produtividade_agricola_no_Brasil_1970-1995_Agricultural_Productivity_in_Brazil_1970-1995/links/0c960537e8798bf85d000000/Produtividade-agricola-no-Brasil-1970-1995-Agricultural-Productivity-in-Brazil-1970-1995.pdf>. Acesso em: 15 de julho de 2017.

VERSCHELDE, M.; D'HAISE, M.; RAYP, G.; VANDAMME, E. Challenging small-scale farming: a non-parametric analysis of the (inverse) relationship between farm productivity and farm size in Burundi. **Journal of Agricultural Economics**, Oxford, v. 64, n. 2, p. 319-342, jul. 2013.

WANG, M.C.; van RYZIN, J. A class of smooth estimators for discrete-distribution. **Biometrika**, Oxford, v. 68, n. 1, p. 301-309, abr. 1981.

WANG, J. **Nonparametric regression analysis of longitudinal data**. 2003. 17 p.

YATCHEW, A. **Semiparametric regression for the applied econometrician**. Cambridge: Cambridge University Press, 2003. 213 p.