On the Feasibility of Data-centric Modeling of Gainesville Businesses[1]

Michael Elliott*, Amardeep Siglani*, Mark Girson*, Erik Bredfeldt+, Lila Stewart+, Matthew Collins*, Renato Figueiredo* and José Fortes* (* University of Florida + City of Gainesville)

I. Introduction and Goals

Cities that thrive economically and provide high-quality of life to their citizens depend on information technology to collect and analyze many kinds of data. The ultimate goal is to understand how city services are meeting their purposes, how private businesses are performing, how civic initiatives are being enabled, how city environments are being protected and, in general, how a city is functioning as an integrated whole that serves its citizenry. Increasingly, the objective is to create smart cities where data collection and analysis enables well informed decisions by city governments and citizens to create metropolitan areas that are safe, economically sustainable, socially harmonious and environmentally friendly. The City of Gainesville (abbreviated as the City) already collects several kinds of data in order to support its administrative duties to enforce regulations, protect citizens and manage services. This document summarizes our investigations of the possibility of analyzing these and other datasets to understand the performance and distribution of businesses in Gainesville. A full report is available from the ACIS Laboratory of the University of Florida.

The Department of Doing of the City is creating mechanisms to support businesses at the different stages of development by providing tools and personal interactions that inform, guide, and enhance efficiency of business actions and operations. We adopted a business lifecycle model inspired by the Blue Ribbon report of the Advisory Committee on Economic Competitiveness to the Gainesville City Commission[2]. This model considers the following three phases of a business (in chronologic order), each with several of a total of thirteen stages:

- Start Phase: includes Dream, Plan, Finance, Legalize and Brand stages during which the Department of Doing helps identify resources and expertise to understand the local market, available locations and spaces.
- Setup Phase: includes Search, Shape and Build stages during which the Department of Doing helps business owners understand planning, zoning, and permitting constraints.
- Operational Phase: includes recurring Hire, Taxes, Celebrate, Open and Grow stages during which the Department of Doing acts as a connector, enabling business owners to identify and source necessary vendors, partners, and services needed to open.

One of the purposes of this study was to investigate whether available data can be used to improve our understanding of how City actions (or lack thereof) impact (positively or negatively) the efficiency and success of businesses throughout their lifecycle. Parts of this broad objective include understanding whether City-business interactions are hindering or facilitating business

[1] This study was supported in part by the Office of the Senior Vice President & Chief Operating Officer of the University of Florida.

[2] http://www.cityofgainesville.org/Portals/0/clerk/CityComm/BlueRibbonReport.pdf

creation and operation and whether the City can gain insights into the reasons for success or failure of businesses. In support of this objective, the approach tested in this project is to use data from multiple sources to infer the efficacy of City-business interactions and identify business health indicators. The goal was not to do an economic analysis of Gainesville businesses, but this approach could support such analysis in the future.

A related purpose of this project was to investigate what data should be collected, how to collect and improve them and how to enable data access and analysis. As additional results, this study includes recommendations for improvements in collecting and managing data to complement efforts by the City, which already provides access to over 250 sets of data related to Economic Development & Redevelopment, Environment & Energy, Governance, Human Potential, Infrastructure & Transportation, Public Safety, and Quality of Life.

Methodology: We started by identifying datasets that are available and potentially could inform us of the nature, distribution, environment and performance of businesses in Gainesville. Datasets were obtained from public government sources at the City, State and Federal levels. We mined these datasets for information of the types of businesses, their geographical distribution, their interactions with the City and factors that could impact business performance. To supplement these sources we conducted a survey of Gainesville businesses to get data on their performance and sentiment regarding their interactions with the City. We then looked for correlations between performance information reported by survey respondents and data from other datasets. Additional analyses were conducted on the datasets in order to discover facts, trends and patterns that could be further used to infer business performance and factors impacting it.

Data availability and sources: The different sources of data include the Gainesville City Open Data web site[3], the City and Gainesville Regional Utilities (GRU), survey data and other sources. They include active businesses (7,044 records), building permit requests by businesses for repairs or expansion in Gainesville (16,795 records), crimes in Gainesville (123,000 records), electricity consumption by consumers in Gainesville (29,359 records), building code violations by businesses in Gainesville (550 records), zoning code violations by businesses in Gainesville (543 records), lifespan of active and inactive businesses (7,088 active business records, 8,651 inactive business records) in Gainesville, utilities consumption of consumers in Gainesville, parcel data for businesses in Alachua County (148,277 records), American Community Survey's 2012-2016 5-year estimate of people living in each tract of Gainesville (57 records), statistical data calculated from answers by Gainesville businesses to 44 questions on business performance and interactions with the City, quantified consumer sentiment regarding the state of the national economy, quantified consumer sentiment of Alachua and Gilchrest county residents regarding the state of the economy, quarterly national Gross Domestic Product measurements by category, yearly city-level Gross Domestic Product measurements by industry and income distributions of census block groups in Gainesville.

[3] Gainesville City Open Data: https://data.cityofgainesville.org/

II. Gainesville City Businesses

According to the data available on November 1st, 2017, there are approximately 5528[4] active businesses in Gainesville. 21.97% of them are classified as retail, 18.58% are professional, scientific, and technical services (including lawyers, attorneys, tax services, auto repair and laundromats) and 18.13% are other services (including barbers, tailors and car cleaning).

The life span of currently active businesses ranges from less than one year to just under 70 years. More than half of these businesses started operating in 2008. 225 businesses have been active for more than 30 years, and only 24 businesses have existed for more than 40 years[5]. From 1999 till 2007, many new businesses were started in Gainesville but a large fraction of these businesses are no longer active. The number of business closures was the highest between 2000 and 2008 (between 400 and 800 business closures per year), as shown in Figure 1.

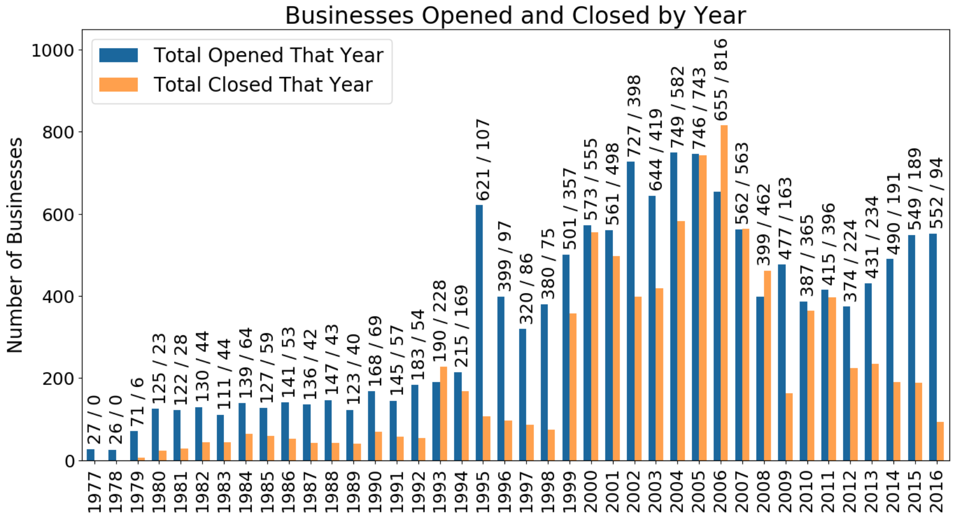

Figure 1 - Blue bars depict the number of businesses that started in a given year. Orange bars depict the number of businesses that ceased operation in a given year.

Geographical distribution of Gainesville businesses: The geographical distribution[6] of the 4,972 active business locations in Gainesville shows that 60% of them (2,968) are located in the NW part of the city (see Figure 2). The SE part of Gainesville is home to only 283 active businesses which corresponds to approximately 6% of the total. In comparison, the population distribution[7] shows that 41% of the Gainesville residents live in the NW while just 3%[8] live in the SE. The

[4] The number of businesses has been inferred from records from different datasets after elimination of records that are likely to be duplicates. Including duplicates would raise the number of active businesses to 7044.

[5] 1995 data (22 years ago in relation to 2017) is inaccurate; record dates reflect record creation dates rather than business start dates; this seems to be an idiosyncrasy of the introduction of new software in the management of data in 1995. This affects data for 1995 in several datasets and plots in this summary and in the full report.

[6] Gainesville was divided into four geographic quadrants defined by the intersection between Main Street and University Avenue. Businesses with multiple entries in the dataset were only counted once per unique location.

[7] Population data were taken from census tract populations provided in the American Community Survey's 2012-2016 five-year estimate. For census tracts that are only partially inside Gainesville's municipal boundaries, the population was estimated according to the fraction of the tract's land that lies within those boundaries.

[8] This number may be low due to the methods we used to estimate population. See previous footnote.

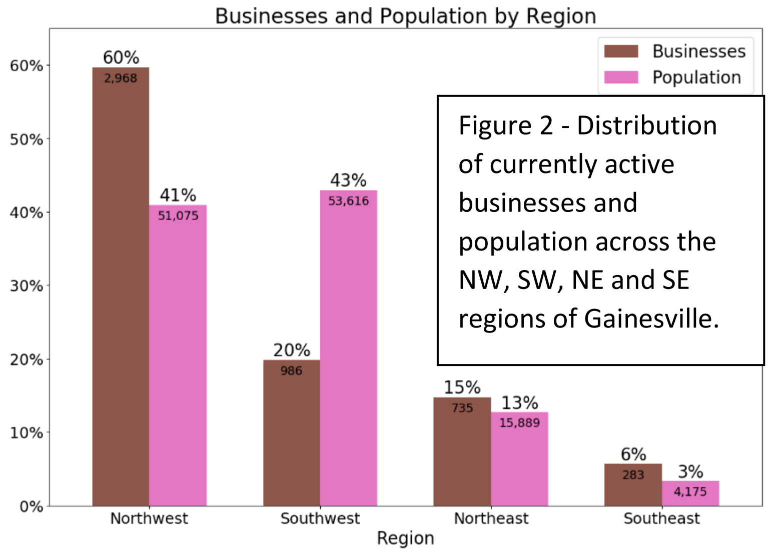

Figure 2 - Distribution of currently active businesses and population across the NW, SW, NE and SE regions of Gainesville.

population-to-business ratios of the NW and NE quadrants are approximately 17 and 15 people per business, respectively, whereas the SW is far more residential (54 people per business) and the SE is only slightly more residential (21 people per business). The SW is the only region that contains a proportion of Gainesville's population (43%) that is larger than its proportion of Gainesville's businesses (20%).

III. Gainesville City Businesses Sentiment

We conducted a survey[9] of the businesses in Gainesville to gauge their sentiment of how successful they are and their satisfaction with their interactions with the City. Of those that started the survey, 479 completed it, 165 partially completed it. The response rate was 16.9%. More than 95% of the respondent businesses have less than 50 employees. More than 65% and 40% started since 2001 and 2011, respectively. When asked about the success of their business, 45.8% answered "very successful" while 46.2% answered "somewhat successful."

The survey asked whether businesses would like the City to provide services during each of the phases of their business life mentioned in the introduction of this report. As shown in Figure 3, for each of the thirteen stages, at least 30% of the respondents responded "Yes".

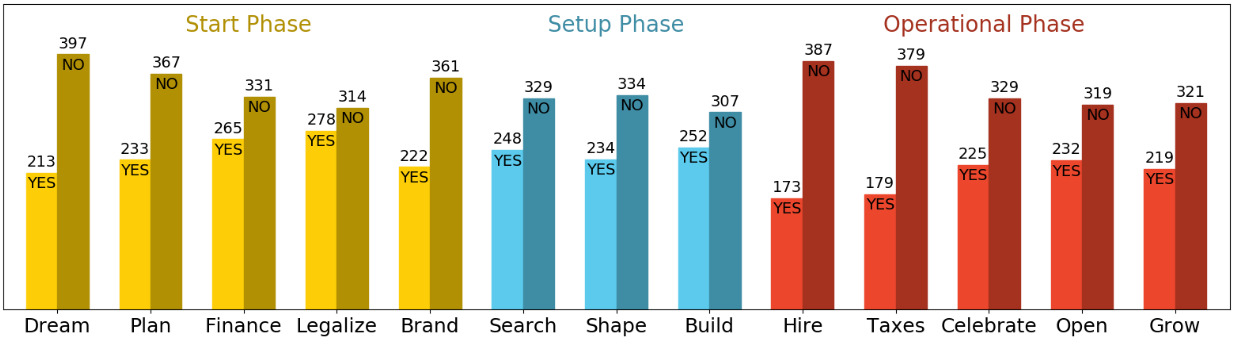

Figure 3 - Numbers of survey respondents who responded Yes or No to the question of whether the City should provide services that support the different stages of a business lifecycle.

The survey also asked each business about its sentiment on interactions they might have had with the City when searching for a business location and/or drawing development or building plans. All departments had an average rate higher than 6 (on a scale of 0 (worst) to 10 (best)).

[9] We used the Active Businesses list available on the City website, with those located outside the city limits removed. On October 17, 2017, an email invitation to take the survey was sent to 4,000 businesses.

IV. Combined Data and Survey Results Analysis

The survey data and the datasets described were combined to find out whether there was a meaningful correlation between publicly available data and the likelihood of business success. The following factors were considered: Gross Domestic Product (GDP), Consumer Sentiment Index (CSI), business age, electricity consumption, crime numbers, building permits, building code violations, and zoning violations.

There is a strong correlation (90%) between the GDP of the US and the GDP of Gainesville. Figure 4 shows that there is also a high correlation (about 82%) between national GDP and business creation in Gainesville. The correlation of national consumer sentiment in a given year and business creation two years earlier is high (0.72).

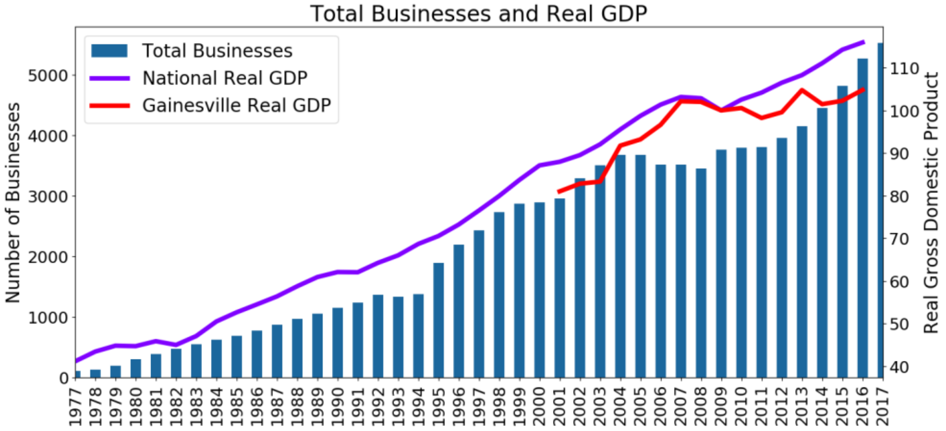

Figure 4 – National and Gainesville MSA gross domestic products (purple and red lines, respectively) alongside the total number of businesses in Gainesville for each year.

We found no correlation between each of the other five factors (electricity consumption, crime numbers, building permits, building code violations, and zoning violations) and business success. Of the forty-four census tracts[10] in Gainesville, all but four have high business success rates above 80%. Three of those exceptions occur around the SE quadrant of Gainesville.

V. Data Management Practices and Analysis Tools

Data availability and accessibility: The City's Open Data portal provides access to a variety of datasets and is a useful resource for data-centric studies of Gainesville businesses. However, these data are insufficient to conduct meaningful analysis. These data, the survey data and additional data should be collected and made more easily accessible. Also needed are documentation of what each dataset contains, how data are formatted, whether vocabularies are used and their logic organization schema.

Data for everyday business operations and their economic performance could help us develop predictive data-centric models. However, regulations that protect private business data make it hard to have access to explicit information on individual business performance. The prospect of indirectly getting business performance information from social networks and web sites also deserves further investigation. However, recent regulations from governments and social media companies might limit access to those data to less than a useful level.

[10] Gainesville was subdivided for spatial analysis using the census tracts defined by the Census Bureau.

Data quality and organization: The quality and completeness of the data available for this study need to be improved if similar data are to be useful for future studies. Numerous records have erroneous entries that cannot be automatically corrected thus requiring laborious human intervention. Data-entry validation mechanisms that check for proper formatting of data and compliance with vocabularies could help mitigate some of the errors found in the data. Checking against other datasets would also be helpful in detecting wrong data. For example, phone numbers, area codes, street addresses and dates are often partially verifiable at data-entry time. In the absence of data-entry validation and best practices, only limited data cleaning might be possible in automated ways.

A major challenge faced by this study is the sparseness of the data. Very few businesses have data in all the different datasets. Thus, it is difficult to give a clear binary answer about the success of a business from the existing data. Another example of incomplete data is the absence of business-type information in most datasets. Another challenge is the presence of duplicate entries in a single dataset. For example, the Active Businesses dataset has multiple entries that differ only in business type. It is hard to automatically correct such entries because they could either be redundant (thus candidates for deletion) or legitimate entries that describe individual entities that are engaged in multiple kinds of businesses. The use of taxonomies or vocabularies is also desirable. For example, business types could be consistently described using North American Industry Classification System (NAICS) codes. This would facilitate data analysis per business type as well as correlation with other datasets that might use the NAICS classification.

Desirably, different data sources should come up with data schemas that use unique business identifiers. This would reduce the efforts required in data gathering and cleaning and allow researchers to devote more time analyzing the data. Data formats need to be agreed upon or, at least, well documented and enforced for each dataset. This applies, for example, to geographical addresses across different datasets so they can be efficiently and accurately matched using software. This will help greatly in merging and analyzing multiple datasets.

Data analytics: We used standard tools and technologies for data analytics and visualization, relying on open-source database and spreadsheet software[11] to merge and query data originally available as spreadsheets. This approach was adequate for the limited number of datasets of rather small sizes considered in this study. For purposes of visualization, we also used open-source tools to generate the graphs and maps[12], as exemplified by the figures included in this report. The tools used in this study were chosen to so that all results are reproducible using the same data.[13] The size and number of datasets for the kinds of analysis needed for this and similar studies are not very large. As long as future datasets contain a negligible number of errors to be manually corrected, the methods used in this analysis are general and can scale to any amount

[11] All data analysis and graphing were completed using the tools provided in the Anaconda Distribution (https://www.anaconda.com/), including Jupyter Notebook (http://jupyter.org/) running Python 2.7.

[12] Map visualizations were completed using the Google Maps JavaScript API (https://cloud.google.com/maps-platform/) running on a Node.js server (https://nodejs.org/en/). Data was housed using MySQL.

[13] These scripts and a snapshot of the public data used are available online at https://github.com/acislab/DGLIM.

of data. Analysis was automated on entire datasets, not done manually. It is possible to mine new types of data that might need advanced analytics on huge amounts of data. Tools and expertise are also available to handle such scenarios.

VI. Conclusions

The aim of this project was to investigate the viability and potential value of using data collected by the City and other public data to identify, quantify, characterize and correlate factors for success during the lifecycle of businesses in Gainesville. We were able to derive interesting observations from the data we collected, including the different types of Gainesville businesses and their distribution, chronological trends of their numbers, annual rates of business creation and closure, geographical distributions of all businesses and correlations between Gainesville business performance and economic indicators, including local and national gross domestic product and consumer sentiment indexes. We also identified factors for which we could not find any relation to business performance. It was not our objective to study the economic performance of Gainesville or its business environment. However, our observations and methods could provide information for such studies from which additional questions could be answered or quantitative evidence could be derived to support qualitative claims.

This study was possible because the City follows an open-data policy that provides to the public access to a large number of City-related data. However, this study also revealed limitations and challenges of data-centric analytics using currently available data. Problems exist regarding the quality, completeness, uniformity, unique identification, representation and logical organization of data from multiple datasets currently maintained by the City and associated entities. These problems limit the extent, generalizability and reliability of studies using available data.

Our main conclusion is the confirmation that there are significant benefits and opportunities to be gained from a City-level data management and analytics framework that enables the collection of high-quality data and cross-dataset queries. The framework should be general enough to also support other types of City-relevant data such as data from environmental sensing, traffic monitoring, public transportation monitors, citizen-provided information, personal devices, City-deployed sensors, operational records from private and public entities who wish to share data. Other cities are exploring how such a framework should look like and be deployed. There is also growing interest from industry and businesses in providing solutions for the whole or parts of the framework. The City should leverage these ongoing developments and explore how they might suit the needs of Gainesville as it becomes a “New American City”.

This project brought together researchers from the Advanced Computing and Information Systems lab and the Bureau of Economic and Business Research at the University of Florida, and the City’s Department of Doing. It provided educational experiences to two graduate students. The team reached out to several branches of the City and Gainesville businesses and shared ideas with the City. Continued encouragement of interactions between UF and City researchers can further contribute to improved City operations and education and research on challenges faced by cities such as Gainesville.