Identifying Venues for Female Commercial Sex Work Using Spatial Analysis of Geocoded Advertisements

Daniil Voloshin, Ivan Derevitskiy, Ksenia Mukhina and Vladislav Karbovskii

1ITMO University, Saint-Petersburg, Russia
achoched@gmail.com, iderevitskiy@gmail.com, mukhinaks@gmail.com, vladislav.k.work@gmail.com

Abstract
Despite being widely visible on the web, Internet-promoted commercial sex work has so far attracted limited attention from the side of researchers. Current studies outline the issues that new forms of sex work are associated with, however, very little is known to date about their spatial manifestation. In this research we follow the environmental perspective in spatial analysis of crime and deviance with the assumption that the location of venues for provision of commercial sex work can be modeled via the algorithms trained on the distribution of possible correlates in the proximity to the existing venues. Visualization of the acquired results is presented herein along with the errors and score metrics for evaluation of the applicability of specific methods of machine learning. The paper is concluded with the estimation of potential extensions and peculiarities of data used in the research.

Keywords: spatial analysis, commercial sex work, machine learning, data analysis, venue location

1 Introduction
Commercial sex work itself and an ever-extending range of accompanying problems pose a constant challenge for policymakers, specialists in public health and law enforcement, society in general and, what is more important, for those exploited in the industry (who are estimated to amount to 4.5 million people [1], [2]). It has not avoided the influence of the development of communicative technologies which has partially shaped its modern image. Given the advantages Internet offers, people involved into commercial sex work in different roles have seemingly minimized their physical visibility in the eyes of law enforcement and public health officials and acquired new techniques of “recruiting” and addressing potential clients with new services more efficiently. Despite the on-going virtualization of commercial sex service, Internet-mediated escort has not yet replaced completely [3] the street-based commercial sex work. Instead, it raised new threats and kept the workers nearly as vulnerable as they were before web services became an important part of so-called shadow economy.
Moreover, it has put to trial the methods and approaches previously incorporated by officials and other stakeholders fighting human trafficking. Provision of services in question has not lost its physical manifestation, but the tasks for analysis of its scope became more complex as the available data became more voluminous and ambiguous. Taking that into account, it is assumed herein that the development of new techniques in studying spatial dimension of commercial sex work is required.

The motivation behind the need for more advanced spatial analysis of commercial sex work stems from a number of considerations. First of all, it might be beneficial for understanding the topologies of networks of sexual contacts used in the research on dissemination of sexually-transmitted infections [4]. Moreover, it can be used to inform the strategies for allocating resources aimed at securing sexual workers and preventing dangerous behavior among both providers of escort services and their clients. From the urban studies perspective, assessment of the venue-choice strategies of commercial sex service may bring new insights into the construction and transformation of the image of certain areas of big cities as well as the role of different factors affecting the emergence of areas characterized by the concentration of facilities of certain type. Finally, spatial analysis of data on commercial sex service venues can provide a basis for the detection of clusters of establishments that are potentially involved into human trafficking and forced labor. Though all of the abovementioned tasks can be approached with the retrospective analysis of data, elaboration of the predictive models is suggested here to account for the changing environmental conditions of commercial sex work.

In the domain of crime time and location analysis (which is now attracting increasing attention in both law enforcement practice and academia [5]), the issues of preventing the solicitation and provision of illegal services and goods have been already addressed [6], however, this topic is still somewhat understudied, compared to the similar investigations of violent and property crime types [7] [8]. Moreover, it can be claimed that spatial mapping and analysis of the web-mediated commercial sex service is still yet to be addressed in the literature. In this paper an extension of the use of spatial analysis techniques and approaches developed in criminology is proposed. Present research follows the logic of the ecological perspective in predictive crime analysis and aims at observing the trends behind the location of venues for commercial sex work based on the influence of the surrounding physical structures and objects (entertainment facilities, police stations, transport, rental flats) that are expected to facilitate or hinder the location of establishments or residence of the individual providers of commercial sex service. The area under investigation has been limited to the administrative boundaries of Saint-Petersburg, Russia. Data on distribution of primary units of analysis (existing commercial sex venues) have been gathered from open access sources – specialized websites that promote sexual service in the city. Areas of potential emergence of new venues are acquired herein through the use of predictive models trained on multiple sources of spatial data.

The rest of the paper is organized as follows: first, a survey of relevant research is presented, with two areas of prime interest: commercial sex work and spatial analysis of crime. Paper then proceeds with the description of the methodology that has been applied to the description and analysis of the geocoded data as well as the discussion of the model that has been developed based on the relations between the variables. Further on, results of the exploratory research are presented. The paper is concluded with the key deliverables of the research along with the proposed directions for future studies.

2 Related works

For a long period of time different forms of commercial sex work, predominantly – prostitution, has been in focus of the researchers from different fields (with certain subjects dominating the research at different times [9]). For instance, in the domain of public health studies the phenomenon of commercial sex has been mostly viewed from the perspective of substance use on health conditions of commercial sex workers [10]. However the major interest towards the investigation of prostitution is
Predicting Venues for Female Commercial Sex Work

D. Voloshin et al

fueled by the role that it plays in the transition of sexually transmitted infections (STI) [11]. A number of studies [12] suggested using sexual contact network analysis in order to outline the risks for the commercial sex workers as well as their clients and sexual partners of both groups. Some of them assumed that geography of the nodes and topology of the networks can have effect on the transmission of the STI's [13], [14]. Another issue that is addressed in studies of commercial sex work is human trafficking [1] and related involvement into prostitution [15]. In the recent years, a growing body of research emerged that addresses the role of Internet in empowering sex trafficking [16]–[19] and moreover – the side-effects that increase the vulnerability of the sexual workers and impunity of their clients.

With the advent of web technologies commercial sex work in its various forms have found new possibilities for both arranging and providing sexual service to the clients via Internet, as well as finding new targets for human trafficking. It has been suggested [13] that this form of information spreading is perceived by the clients as a more effective and safe form of interaction with other clients and providers of sexual service. It has been as well noted [13] that the number of sex service providers does not grow linearly with the population of the city under investigation. In this research this observation is tested in a different manner through the analysis of the relation between the local residential population densities inside the city and the “hot-spots” of escort provision. It is also worth noting that there is currently a lack of research that focus on the investigation of the web-based prostitution on the intra-city level, since the majority of research deals with data on scales larger than a single city.

Though internet-mediated prostitution is generally perceived as being a domain of individual commercial sex workers practicing independently from brothels and other establishments of the kind, there is a growing body of research that seeks to detect the cases of human trafficking in the overall mass of escort web advertising. For instance, in the approach proposed by [20] the opportunity for human traffickers, provided by the classified advertising services (such as Craigslist.com and Backdoor.com) is used for the benefits of law enforcers. Analysis of the full-text contents of advertisements (through the identification of common expressions patterns) and meta—data is assumed to be useful in identifying potential cases of forced involvement into prostitution and patterns of traffickers’ relocations. Other researchers [21] take the idea even further and propose to use the data from classified advertisements to identify potential traffickers and, provoking the artificial denial of service of the telephone numbers listed in the entries, make such web advertising useless [21]. A more systematic approach is being taken by [22] and [23] who propose to use the machine-learning algorithms, informed by the practitioners in the sphere of fighting human trafficking, to analyze the textual contents of the advertising and adjust the models that would effectively separate the cases of individual provision of sexual service from the forced involvement into the sex trade. One of the main benefits that these approaches present is the use of so-called weak features in order to classify the entries other than relying solely on strong features, such as telephone numbers that appear to be dynamic. Despite these approaches showing plausible results, there is a significant limitation that hinders their wide application – their dependence on the availability of large amounts of reasonably consistent data presented in a textual form. This condition is especially sensitive if one considers the cases of specialized websites such as the ones analyzed in the present research – which store the data in partially closed format with minimum verbal descriptions attached.

Considering the spatial analysis of commercial sex work, it is worth noting that the majority of studies to date have been focused on describing the localities of street-based [24], [25] and establishment-based escort [26], [27] and thus seem to have neglected the possibility of investigating the ways web-based commercial sex work is materialized in the urban space and what are the areas of intersection between the two of these types. In order to fill this gap, present study suggests exploring the possibility of setting up a model based on the open data from web escort advertising websites for identifying potential locations of new commercial sex venues.
Since commercial sex work is both a complex social phenomenon and an area of occupation, which is stigmatized in the majority of the modern societies, for a long time it has been of prime interest for practitioners and researchers from the fields related to public health, social service and law enforcement. In the latter domain a number of techniques have emerged that assist the analysis of the scope and spatial patterns of the problem on a regional or city level. Spatial analysis of crime has been initially motivated by the practitioners and policy makers who sought for mechanisms of sex service solicitation prevention. For quite a long time advances in both theory and practice were focused on understanding the background, motivation and behavioral patterns of individual offenders [28]. Along with the latter the interest has been as well paid towards the understanding of the social embeddedness of crime, its ethnicity, socio-economic and other features of communities [29], [30]. When it became obvious that crime events are distributed unequally through the space [31], practitioners and researchers started developing concept-backed methodologies that would allow the detection of areas, where most of the law violations happen.

Over the course of time, a number of techniques have been developed with some of them successfully engaged into the everyday practice of police departments. The most common way to categorize them found in the literature [7] is to divide them into retrospective (also referred to as “hot spot mapping”) analysis official crime records (which incorporates a wide range of techniques [8]) and more sophisticated, theory-backed analytical methodologies like Risk Terrain Modeling [32]. The difference between the two approaches is that hot spot mapping derives predictions for future occurrences of crime based on the locations and congregations of already-registered law violations, thus hypothesizing that crimes are likely to happen in those areas where cases of greater incidence frequency or quantity are concentrated [28]. Risk Terrain Modelling extends the application of geocoded official crime records with an overlay of the data on spatial distribution of the factors that have been empirically proved to correlate with an increase or decrease in the risk of the incidence of new crime events of a certain type (residential burglary, car theft, aggravated assault etc.). As a result of applying such a technique an ad-hoc map (produced for the particular spatial unit of analysis) representing the areas of relatively higher or lower risk is acquired and used to inform future crime mitigation operations.

Though present research has been significantly inspired by Risk Terrain Modeling and partially by its predecessor, there is a number of limitations that do not allow a straightforward application of the above-mentioned techniques currently developed in criminology. First of all, both approaches require substantial history of records of actual crimes in order to make predictions reasonable and statistically significant. In this sense, access to the reliable and systematic data on crime is required, which is to times exclusive for the researchers affiliated with law enforcement agencies or their partners. Secondly, as it has been mentioned above, the empirical and conceptual basis behind the prediction of future locations of internet-mediated commercial sex work is now limited. In order to address this issue, it is proposed to explore first the relevance of the correlates of the street-based commercial sex work found in the literature [32] as well as the ones suggested in this paper given their alleged link to the elements of infrastructure purportedly required for the promotion and provision of commercial sex services (referred to as opportunities in the field of crime studies [33]). Along with the attractors and generators of the observed type of illegal activities, the role of the mitigating factors in “repelling” commercial sex venues (the presence of the police stations in the nearest proximity) is assessed as well.

3 Dataset description

Geoinformation systems (GIS), which are proclaimed to be a perspective technology to assist policy-making in the field of crime prevention [34] are as well used by those people standing behind the internet-mediated commercial sex trade (whether they are managers or individual providers of
commercial sex services) and the final consumers of the service – the clients of escort agencies, brothels or individual sex workers. Though this factor facilitates the expansion of the escort economy, it nevertheless provides researchers with a flexible source of easily quantifiable data.

The following is the description of the datasets used for mapping the approximate locations of the venues for commercial sex service provision and distribution of the points marking the presence of the environmental factors that are expected to correlate with them. The inclusion of a primary set of factors has been motivated by assumptions derived from the existing literature on the spatial correlates of another type of commercial sex work – street prostitution [35]–[37]. Two types of datasets for the case of Saint Petersburg have been collected and analyzed:

1. Independent variable-related: originally the dataset included point locations of bars, saunas, objects of so-called adult industry (entertainment facilities that might potentially attract major flows of clients), police stations (spatial limiters of illegal activities including commercial sex trade), metro stations (denoting accessibility of the regions where commercial sex venues might cluster) and hotel or apartment accommodation offers (suggested herein as an important part of the commercial sex infrastructure). Further on, it was extended with additional categories in order to explore a wider range of latent correlates: banks, tourist attractions, shops, parking lots, beauty salons, industrial sites, etc. Moreover, point data have been supplemented with municipality-level population density and inequality statistics. The data on differences in well-being among residents of municipalities has been derived from the study of spatial characteristics of socioeconomic inequality in St. Petersburg [38]. Author assessed well-being through weighted sum of the following attributes: property taxation rates, self-employed to working population share, higher education attainment and the proportion of residents holding a scientific degree.

2. Dependent variable-related: geocoded advertisements data collected from websites promoting commercial sex services via HTML parsing algorithm adjusted to gather entries’ metadata.

Locations of the objects from the first dataset have been derived from the open-sourced Wikimapia geographic information service based on participant contributions. Coordinates for each of the objects from the described categories have been retrieved via the API provided by Wikimapia. Throughout the study, it has been noted that entries in the database are not always accurate. Most of the errors can be attributed to the nature of the geocoded information in the database as it has been submitted by the users on the voluntary basis. For instance, a multitude of points tagged as “bars” and located inside certain building may equally mean that there is one establishment of this type found at this address or a number of them. Due to the same reasons, some of the new objects appear in the database with a delay.

Location of police stations have been derived from the official website of the regional Internal Affairs Office [39], and transferred into geographic coordinates. Dataset comprising the locations of subway stations includes geographical coordinates of the points matching the center of the lobby of each of the stations of the St. Petersburg metro.

Rental estate data has been collected from the local classified advertisements website [40]. The total count for apartments offered is 10000 for Saint-Petersburg. The priority has been set for those premises that have one or two rooms and studio-type flats. The shares are the following: a total of 959 advertisements for studio-type apartments, 6294 for a single-roomed flats and 3587 entries offering two-
roomed accommodation. The data has been retrieved with the help of the algorithm that processes the GET-requests to the website. The output consists of the coordinates for each of the addresses acquired with the help of the geocoding algorithm described above and a price set for the rent of a particular object (either per day or per month depending on the type of advertisement).

Commercial Sex website data collection has been organized in the following steps:

1. Selection of websites that present geocoded profiles of potential commercial sex workers. The primary inclusion criterion is the availability of geotagged profiles of potential sex service providers. In order to collect a sufficient number of cases, only the websites with at least 100 geocoded profiles have been chosen. In addition, during the visual analysis of the geotags, a number of websites was excluded from the analysis because of the alleged use of the algorithm for subsequent arrangement of components in a regular grid, which proves the artificial nature of these spots’ location. Figure 1 presents an example of one of such cases.

2. Acquisition of the coordinates, price parameters and phone numbers (used for filtering out the cases where same phone numbers are used on different websites) for each of the profiles. A total of 2100 entries have been collected at this stage.

Exclusion of the repeating entries is based on the matching mobile phone numbers between the websites under investigation. Since in the present study we do not aim to address the issue of separating brothels from individual escort providers, in cases where two profiles with identical numbers are detected on the same site, both profiles are added to the dataset. The resulting dataset consists of 1796 geo-coded sex workers’ profiles.

4 Experimental study

4.1 Methodology

Given the multiplicity of factors that could be potentially associated with the distribution of commercial sex venues, analysis of the data have been performed in two steps: (a) calculation of p-values (used to filter out insignificant spatial factors) and Pearson-R coefficient (aimed at distinguishing the role of factors most associated with the allocation of commercial sex venues); (b) application of machine learning algorithms to the development of the model for commercial sex venue identification.

The primary unit of analysis (location of a commercial sex venue) is represented by a point on a grid with a cell size equal to: (a) area of municipality for irregular grid (used in correlation analysis); (b) 1 square kilometer on a regular grid (used in modelling). The size of the cell for a regular grid has been adjusted following the visual analysis of the results of the test runs as to maintain the balance between accuracy and precision. The use of different grids for correlation and regression analysis

![Figure 2: Heat maps produced by different machine learning algorithms (a-g) and the one acquired from the visualization of the primary data (h)](image-url)
is motivated by the need to achieve results that could be meaningful regarding the existing administrative division of the city and allow for their higher reliability.

Since there are currently no empirically-grounded estimates of the effect that various environmental factors have upon the internet-mediated commercial sex venues location, the aim of identifying new venues has been approached through the regression analysis using the following machine learning algorithms (for a comprehensive review of the models, refer to [41]): (a) Lasso; (b) Linear Regression; (c) Random Forest Regressor; (d) KNeighbors Regressor; (e) Support Vector Regressor; (f) Bayesian Ridge; (g) Gradient Boosting Regressor.

Two types of features associated with these objects have been introduced into the development of predictive models:

1. Radius-based – stand for the number of objects found in an area around the given point. The radius of this area for points representing independent variables has been set up to 2 kilometers, in order to account for potential overlapping effects of the nearest points of interest.
2. Range-based – proximity to the nearest objects of certain type.

The target value for predictive models is a number of suggested venues (marked by the individual advertisement entries) that fall into the range of 1.5 kilometers around the spot where an object, identified as having certain effect on the likelihood of emergence of new venues for commercial sex service.

### 4.2 Results

At the primary stage of the data analysis, a number of categories of objects have been excluded from consideration on the basis of computed p-values (p > 0.05): markets, educational facilities, industrial areas, nightclubs, striptease clubs and saunas. The linear correlation coefficient has been calculated for the following categories: (Fig.3): (1) Hotels, (2) Hostels, (3) Parking lots, (4) Beauty Salons, (5) Sex shops, (6) Population density, (7) Police stations, (8) Bars and clubs, (9) Well-being level, (10) Pharmacy shops, (11) Bars, (12) Metro stations, (13) Two room apartments for rent, (14) Short-term apartment rentals, (15) Shops, (16) Rooms for rent, (17) Long-term apartment rentals, (18) Banks. Correlation coefficients analysis shows that some of the originally assumed environmental factors (metro stations, short-term apartments for rent, leisure spots) have proven to be moderately associated with the commercial sex venue locations, whereas banks (one of the categories that have been included into the extended dataset for the matter of exploratory inquiry), shops and long-term rentals exhibit surprisingly strong correlation with sex trade venues.

![Correlations](image-url)

**Figure 3:** Pearson-R coefficient values describing the relationship between various environmental factors and distribution of commercial sex venues among municipalities
The following are the results of data analysis carried out in pursuit of the goal of predicting the locations of the commercial sex venues based on the characteristics of the neighboring transport, entertainment and accommodation facilities.

In order to estimate the applicability of each of the proposed methods, two tests have been carried out: the visual analysis of the heat maps produced by the resulting models (Fig. 2) and assessment of error levels (Fig. 4). The overall picture generated by each of the methods reproduces the visuals derived from the data quite realistically hence we assume that the proposed methodology can be applied to identifying the already-existing clusters and future locations of commercial sex venues.

From Fig. 3 that depicts the results of the trials (calculation of the mean and median absolute errors, R² scores and explained variance), it can be inferred that Random Forest Regressor (3), KNeighbors Regressor (4) and GradientBoosting Regressor (7) produce the most plausible results compared to other algorithms, given the relatively higher scores and lower error levels.

5 Discussion and conclusion

Though the results are somewhat tenable and suggesting the potential of the proposed approach to identification and modeling of the venues for web-mediated commercial sex, a number of considerations have to be highlighted with regard to the observed associations between spatial factors and moderate, yet higher than expected values for errors generated by the acquired models.

Firstly, it is worth noting that in this research we have modeled the environmental risks of emergence of commercial sex venues in different areas of the city and not the exact location and quantifiable probability of their emergence. In other words, the results of the modeling assume that spatial regions, where the risk is higher, are just relatively more suitable for setting up a venue than those where it is lower. Thus the outcomes shall be best treated as insights into under-investigated domain and proof that it can be accessed with the suggested approach.

Throughout the analysis of the data, it has been assumed that the open-sourced, user-contributed geographical data may not be robust enough – it is rich, but at the same time prone to errors, bias and may lack precision. The distribution of spots is far from being ideally consistent which suggests that the set is prone to incompleteness and existence of “blind spots” observed in different parts of the city. There are also certain concerns about the overall quality of the data on the locations of existing commercial sex venues. Since it has been submitted to open access by people who mediate or organize the provision of commercial sex services, without further analysis, one can not be sure that it is not purely artificial (namely, representing a product of deliberate work of specialists affiliated with sex trade coordinators), based on the promotional patterns that might exist on the market. In other words, further effort shall be put into making sense of the presented data and understanding the underlying mechanisms of Internet-mediated sex trade.
Though the assumptions behind some of the proposed features of the physical environment of the venues have been partially justified with relevant tests, a number of spatial factors have been excluded from the analysis due to the absence of appropriate data. On the other hand, some unsuspected factors (e.g., location of banks, long-term apartment rentals and shops that are unlikely to explain the concentration of sex trade venues) have been found to be strongly associated with distribution of commercial sex venues, which implies the presence of confounding variables that are to be identified in further research.

In order to overcome the abovementioned limitations and expand the scope of the present research, we propose a number of extensions to the data and methodology elaborated herein. A cooperation with law enforcement authorities in accessing crime records could be preferred for objectively assessing the scope and validity of the research. Throughout the analysis of the primary data on venues locations it has been hypothesized that currently available advertisements found on the Internet might be used as an additional source of information for predictions, but before that they must undergo the process of classification in order to separate the locations of the venues where different forms of commercial sex service are provided (for instance, brothels, domestic, street and individual venues). Moreover, as a potential solution to the issue of latent artificiality of the data, we suggest supplementing the analysis of the metadata of commercial sex advertisements with the study of their contents (e.g., full text or closed-format descriptions analyzed under a mixed-method research design). This can help achieve at least two goals: (a) identification of the semantically-linked clusters of venues (assuming the presence of sex trade organizers or mediators); (b) analysis of the distinct features of the occupation-specific language (which can be further used to track the spread of related information through social networking websites [42], one of the important mediums for communicating potential clients). Furthermore, we propose herein that more flexible (compared to the administrative one) classification of regions (for instance, vernacular areas) as areal units of analysis is preferable, as the borders of municipalities can doubtfully match those of congregations of sex trade venues. Finally, we suggest further investigation of the role that transport plays in the allocation of commercial sex venues (for instance, through empirically-grounded agent-based simulations [43]), since transport accessibility of venues is expected to boost their competitive advantages.

Acknowledgements. This paper is financially supported by Ministry of Education and Science of the Russian Federation, Agreement #14.584.21.0015 (11.11.2015), project ID: RFMEFI58415X0015.

References

Predicting Venues for Female Commercial Sex Work

D. Voloshin et al


