MNRAS 000, 1–5 (2020) Preprint 1 December 2020 Compiled using MNRAS LATEX style file v3.0

ProjectX – Infectious Respiratory Diseases and Air Quality

William Glazer Cavanagh[1], Guillaume Quenneville-Himbeault[1], Kenza Benmansour[1], Leslie Moranta[1], Laurent Bruneau-Cossette[1]
[1]*Université de Montréal, Montréal, Québec, Canada*

30 November 2020

ABSTRACT

Air quality is an important factor to human health. Furthermore, it has been established that several air pollutants have an impact on the occurrence and outcome of respiratory diseases. Thus, in light of the current pandemic, it is important to know whether air quality has an effect on COVID-19 propagation. As long-term exposure to particulate matter has already been linked to increased mortality, we set ourselves on expanding the results of [3] on the short-term effect of air quality, temperature and humidity on the propagation of COVID-19.

In this paper, we aggregates and remapped the daily statistics collected on air quality weather models to COVID-19 active cases and mobility data reported by urban counties in the United States of America. We then classified the growth of new COVID-19 cases for a 7-day period in 3 categories: accelerating, stagnating and decelerating. The growth rate was obtained using the second derivative of daily new cases by day obtained with a Savitzky-Golay 7-day cubic filter. We finally predicted the aforementioned growth rate category for each 7-day period using weather, air quality and mobility data from 3 to 10 days prior the beginning of the period. The results show that mobility data is a worse indicator of COVID-19 propagation than weather and air quality data. However, both measures are somewhat poor indicators of COVID-19 propagation overall and exhibit strong regional specificity, with accuracies ranging from barely above random (0.34) to 0.54, depending on the regions of the USA on which predictions were attempted.

Key words: air quality – infectious respiratory diseases – climate change

1 INTRODUCTION AND BACKGROUND

Air quality is an important factor to human health. It has been established that several air pollutants such as NO_2, SO_2, O_3, PM_{10} and $PM_{2.5}$ have an impact on the occurrence and outcome of respiratory diseases [9], [5]. In light of the current pandemic, it is thus important to know whether the COVID-19 pandemic could be related to air quality. Some tentative results have already been obtained that correlate physical properties of ambient air with COVID-19 transmission [8], and recently, a study linked long term exposure to higher $PM_{2.5}$ concentrations with a higher death rate from COVID-19 [7]. Other such studies include [3] which found a positive correlation between $PM_{2.5}$ air pollutants and COVID-19 transmission as well as a negative correlation between temperature and humidity variables and the virus' transmission in the Lombardy region (Italy).

These links having already been established, it remains to be known whether short-term fluctuations in air quality have an effect on COVID-19 transmission and outcome using a larger dataset. As the USA has become one of the countries most affected by the disease, due to the extensive, county-level reporting of COVID-19 cases, it has also become a comprehensive dataset of COVID-19 infections and deaths. Furthermore, due to the fact that case reporting is standardized by the same agency, the number of reported cases by county all use the same methodology. We therefore do not have to adjust for varying reporting practices, and can assume a more uniform dataset. The goal of this study if to expand of the results from [3] on the dataset of COVID-19 infections in the USA [6] using global air quality data from the Copernicus Atmosphere Monitoring Service (CAMS) Near-Real-Time air quality model mapped to counties.

2 METHODOLOGY

2.1 Description of our dataset

We began by listing the primary factors that had the biggest impact on COVID-19 infections rate. Based on the literature ([2] and [4]), the two most prominent factors are the population density and the government regulations put in place to fight the pandemic. With this information, we decided to create our own datasets that grouped subsets of the counties in the USA that have similar population density as well as similar laws and measures in place regarding COVID-19. These subsets are used to isolate the biases that could lead to a wrong correlation between air quality and COVID-19 infection rates. With the goal being to link air quality data and COVID-19 data, there were many external factors which guided our approach. The biggest one being the quantity and quality of available data. This was a major factor why we chose to study the relationship between COVID-19 and air quality instead of some other respiratory disease. We consulted Google Cloud's [1]COVID-19 open data initiative for well curated and up to date datasets. Having up to date data was very

[1] https://console.cloud.google.com/marketplace/browse?filter=solution-type:dataset&filter=category:covid19

important since one of the goals of the project was to be able to give real time analytics. As for air quality, there were many datasets to choose from. We settled on the CAMS Near-Real-Time due to its completeness, ease of access, and familiarity within the team.

Our COVID-19 data came from the COVID-19 Open Dataset hosted by Google [6]. We constrained our preliminary results to the USA, which in this dataset is subdivised by counties. The data included cumulative cases and deaths as well as a mobility score. The mobility score was calculated by Google using Google maps tracking. The metric used the median time spent doing tasks in a county when COVID-19 was not in effect and represented the percent difference from the regular median for that day. More info can be found [2]on the Google COVID-19 mobility website. The goal in using the mobility score for the model was to control for the difference in reaction by the county governments. As expected, there is a very clear inverse correlation between the mobility workplace score (which is the amount of time spent at a workplace) and the county services being shut down (the workplaces). Taking this into consideration for our model will help us reduce the noise caused by the difference in government interventions between counties.

The merging of the COVID-19 dataset with the CAMS data posed some challenges. The major hurdle was mapping air pollution and meteorological data from the CAMS data onto the US census divisions (e.g. by remapping information from regular weather model grids to irregular census divisions) since CAMS data is not county-shaped, but grid-shaped, with a 0.4° by 0.4° grid. We mapped counties and the grid data using a simple weighted average, with weights corresponding to the ratio between the area of the grid square intersecting with the county shape and the area of the county shape. This was done in order to make the data for a given period of time match the geographical regions of COVID-19 cases and deaths which are reported by census divisions in the USA. When this was successfully done, we then had access to several air quality metrics, such as $PM_{2.5}$, PM_{10}, the AOD_{550} of all particulate matter and several of its sub-components (e.g. black carbon, organic matter) and total column concentrations for several pollutants (e.g. peroxyacetyl nitrate, formaldehyde), for the specific county. Although some of these measurements are aggregate measures for all atmospheric levels, they were included in the model, and will all be inspected as either sources of noise or plausible ground-level pollutants with an impact on COVID-19 infectiousness.

2.2 Data preprocessing

We started our process by creating a subset of urban counties. This selection was done using two criteria: at least 2500 people needed to live in the county and the density needed to be of at least 500 people per square mile, in a way [3]similar to the USA Census Bureau criteri. For population and population density, we used the Census Bureau data [4] (mapped to county shapes by ArcGIS).

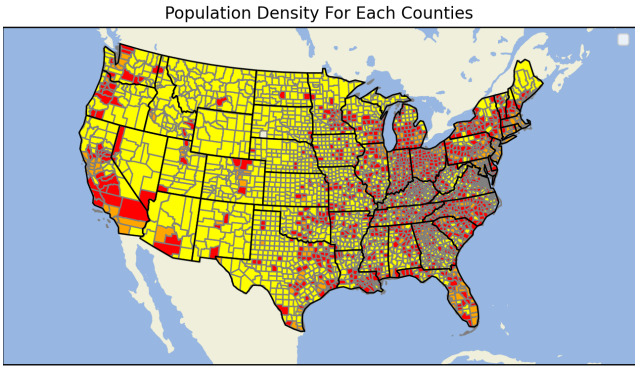

Figure 1: Map of population density by county

With regards to mitigating COVID-19 policy impact, it was harder to account for the change in growth rate due to the laws which were put in place in each county. We first had to choose broad categories to fit laws that contain many different subtleties. Since mask wearing has been proven to have a major effect in the infection rates of COVID-19, [2], we decided that we would base our classification by laws regarding mandatory face masks. As such laws are state matters, we associated each county to their respective state's policy. We then separated states by their mask requirements. The intention was to further isolate the effect of local government intervention so that we could analyze the effect of air quality. The states in the map below in green are those with Mandatory Mask Law for the Entire State and those in red are for the states that have no such law.

In the end, the following subdivisions were made :

(i) The West Coast, containing California, Oregon, Washington and Nevada,

(ii) The East Coast with mask, containing all non-landlocked east coast states with a mask mandate, as well as Washington D.C., Vermont and Pennsylvania, due to their proximity with the state of New York.

(iii) The South with mask, containing Alabama, Arkansas, Colorado, Louisiana, New Mexico and Texas

(iv) The East Coast without mask, containing South Carolina, Georgia and Florida

(v) The Middle with mask, containing Ohio, Michigan, Illinois, Wisconsin and several others

(vi) The Middle without mask, containing most of other states

Complete enumeration of states in each category, by their state FIPS code, is available in the source code.

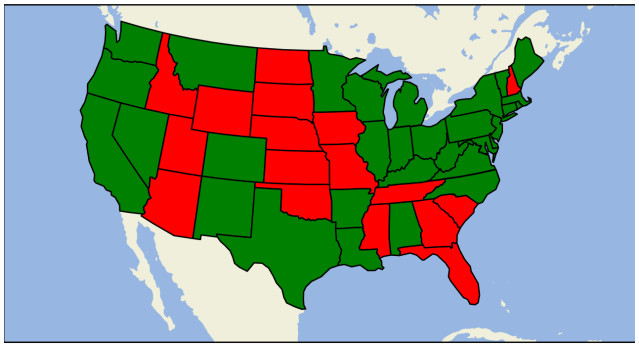

Figure 2: **States with mandatory mask laws**

Regarding the weather and air quality data, approximate measures of particulate matter subcomponents were mapped to their respective proportions, and all values were rescaled and normalized, after which

[2] https://www.google.com/covid19/mobility/
[3] https://www2.census.gov/geo/pdfs/reference/GARM/Ch12GARM.pdf
[4] https://hub.arcgis.com/datasets/fab7849b55d54f0f8f246605f6ee9306/data?geometry=-85.431%2C-0.072%2C-159.962%2C76.663&layer=5

a Principal Component Analysis (PCA) was performed in order to reduce the number of features. Of the 24 available variables, we found the optimal reduction to be of 10 dimensions, which accounted for 94% of the total variance. To further reduce the amount of data, the CAMS dataset which was initially given on an hourly basis was aggregated per day using the daily average, minimum and maximum.

Since the reporting of cases and death for each county were not continuous, some remodeling of the COVID-19 data had to be done. A Savitzky-Golay (or LOESS) smoothing algorithm [1] with a polynomial of degree 3 using 7 points (3 points surrounding the central point from each side) was used to do so. This allowed the preservation of fast fluctuations in COVID cases while also damping noise in these fluctuations. In the graph below, we can see the smoothing for Rockland County in the State of New York (FIPS: 36087). The blue line behind the orange one represents the raw data, and the orange line represents the smoothing using the Savitzky-Golay filter. As the filter smooths the main curve, it can also calculate the first (green line) and second (red line) derivatives of the data according to the polynomial fit. The purple line is a reference for $y = 0$.

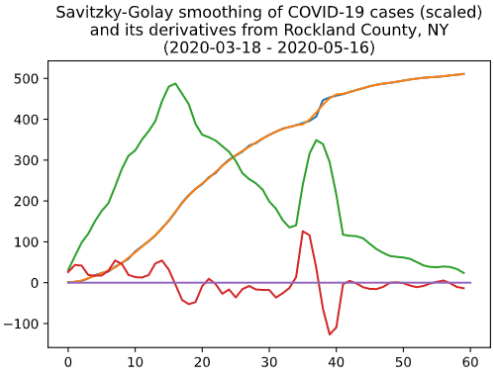

Figure 3: **Effect of smoothing pre-processing**

To make sure we could get more accurate results, we decided to only factor in data with at least 25 new cases a day for consecutive periods of 14 days or more. This further reduces the noise to signal ratio of this data. The resulting 7-day periods periods were categorized into accelerating, stagnating and decelerating trends. The thresholds were set for each category at a variation of 6 new cases per day in the growth rate, thus of their second derivative falling above 6, between -6 and 6 and below -6, respectively. This roughly corresponds to the 25th and 75th percentiles of the second derivatives. The probability distributions for these categories were then fitted against typical distributions for each category to determine the final class in which the given period falls.

2.3 Data analysis

Given the set of mobility and air quality data for 7 consecutive days and the category of the corresponding period, a balanced dataset was created by undersampling the accelerating and stagnating categories to match the number of data points in the decelerating class. The resulting dataset was then analyzed using several standard classifiers to identify the best ones given by either the mobility data, the air quality data, or both. It was found that KNN (specifically 5-NN) and Gradient Boosting Classifiers were the best at classifying the data, with weather and air quality data being significantly better than mobility data at predicting the COVID-19 case trend class.

To verify whether weather and air quality data were sufficient to encode the information on data point geographical location, a cross-validation was subsequently performed on the previously outlined geographical regions of the USA in two ways. First, using the data from a given region, to predict the trends for the rest of the USA, and then doing the opposite, i.e. using the rest of the USA to predict the given region.

3 RESULTS

3.1 Exploratory Data Analysis

Once the final dataset was chosen, we performed an exploratory data analysis. We noticed that the growth of cumulative cases was not the same for every county. While they all exhibited exponential growth, the characteristics of the curves were not the same. Some curves were translated while other had a different level of "flatness". This can be clearly seen in the top 10 most active counties (in #of days with > 25 new cases) .

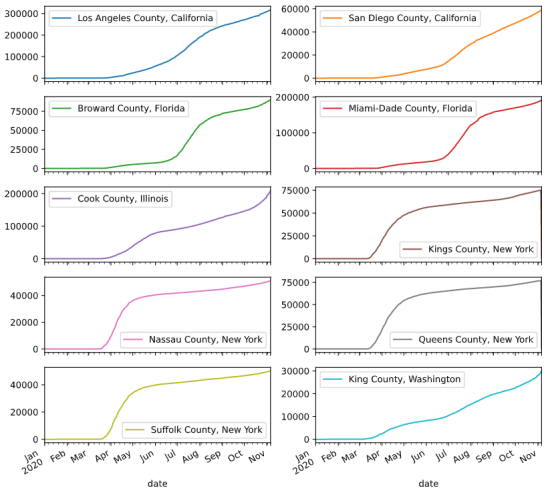

Figure 4: **COVID-19 trends for counties with the most activity**

From these graphs, we see that counties on the east coast saw a dramatic rise near the start of the pandemic until they were able to flatten the curve. The west coast counties started later and were better prepared.

As for mobility, a low mobility score was inversely correlated with people going to work, transit, and groceries. We used this result as a sanity check for the accuracy of Google's mobility score. One interesting data point is that there is no correlation between park usage and the other mobility scores. We chose to include the mobility data in our training to represent the current COVID-19 restrictions in effect. While it is not a direct relation, it does represent it fairly accurately and it avoids the issue of categorizing the type and manner of restrictions imposed on a county.

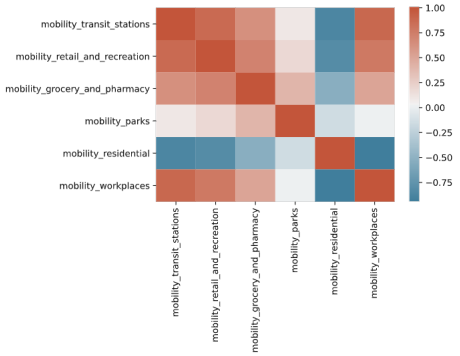

Figure 5: **Correlation matrix of mobility data**

3.2 Data classification

The CAMS, COVID-19 and mobility data were merged using a window of 7 days each, with a 10-day interval between the beginning of the CAMS and mobility data.

Once the two datasets were successfully cleaned and merged we ran a battery of algorithms to see if any were immediately promising. We chose the subset of urban counties with significant COVID-19 transmission (#of days with > 25 new cases). This combined with the fact that we were training on windows of 7 days reduced our training set to approximately 8000 samples, once the dataset was balanced. While it was not a lot, it did allow us to rapidly train and test several models using the scikit-learn python library. We chose the accuracy of each model as a measure of the success of classification.

Each model was first tried with all regions using 10 fold cross-validation. Specifically, we tried gradient boosting, Ada boosting, decision trees and KNN classifiers as well as logistic regression. The best models were the Gradient Boosting Classifier and the k-Nearest Neighbors classifier (with 5 neighbors).

The best two models were the 5-NN on CAMS (mean, min or max), with an accuracy of 0.597, and the Gradient Boosting classifier on mobility and CAMS data, with an accuracy of 0.585 (Table 1). This is significantly better than random, meaning either that CAMS data can be used to predict, with limited accuracy, the trend of COVID-19 propagation, or that the geographical location of the data points is encoded in the CAMS data.

To determine whether location was learned indirectly through the data, the dataset was split in six different regions, and this split was used as a way to perform cross-validation. This specific split yields much worse results than a random splits (Table 2) with 0.41 ± 0.03 accuracy for the 5-NN classifier and 0.40 ± 0.03 for the Gradient Boosting Classifier, indicating that geographic location is indeed encoded in the CAMS data, thus meaning that learning an efficient classification of COVID-19 trends is mostly a matter of location.

What is interesting, however, is that although the models that were tried performed poorly, they still managed to avoid some of the confusion between accelerating and decelerating trends. Indeed, the most probable confusions happen between accelerating and stagnating, or stagnating and decelerating, but not between accelerating and decelerating. This means that although air quality and weather data are poor indicators of COVID-19 trends by themselves, they do hold some meaningful information regarding these trends.

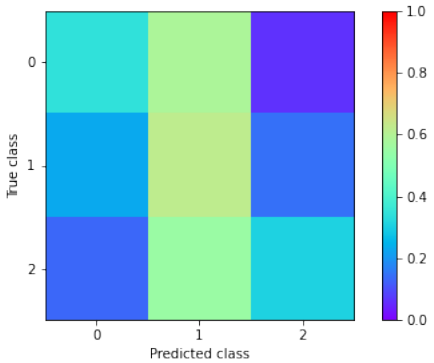

Figure 6: **Confusion matrix of prediction using the East Coast with mask as a reference.** The confusion between accelerating and decelerating trends is lower than the confusion between other combinations of different categories.

Model	Accuracy	Std. Dev.
Mobility		
Gradient Boosting	0.522	0.02
Ada Boosting	0.465	0.021
Decision Trees	0.446	0.014
K-Nearest Neighbours	0.478	0.011
Logisitic Regression	0.434	0.017
Mobility + CAMS (mean)		
Gradient Boosting	0.585	0.022
Ada Boosting	0.524	0.017
Decision Trees	0.504	0.022
K-Nearest Neighbours	0.482	0.025
Logisitic Regression	0.524	0.022
CAMS only		
Gradient Boosting	0.571	0.017
Ada Boosting	0.499	0.016
Decision Trees	0.492	0.013
K-Nearest Neighbours	0.593	0.012
Logisitic Regression	0.5	0.011
CAMS only (first 6 PCs)		
Gradient Boosting	0.553	0.023
Ada Boosting	0.49	0.019
Decision Trees	0.479	0.015
K-Nearest Neighbours	0.576	0.013
Logisitic Regression	0.503	0.011
CAMS only (max only)		
Gradient Boosting	0.573	0.017
Ada Boosting	0.505	0.016
Decision Trees	0.491	0.028
K-Nearest Neighbours	0.591	0.014
Logisitic Regression	0.504	0.018
CAMS only (min only)		
Gradient Boosting	0.558	0.013
Ada Boosting	0.5	0.014
Decision Trees	0.493	0.012
K-Nearest Neighbours	0.589	0.02
Logisitic Regression	0.506	0.018
CAMS only (mean only)		
Gradient Boosting	0.559	0.018
Ada Boosting	0.507	0.014
Decision Trees	0.496	0.027
K-Nearest Neighbours	0.597	0.015
Logisitic Regression	0.512	0.019

Table 1: Preliminary results on whole dataset

		K-Nearest Neighbours		Gradient Boosting	
Region	**Policy**	**Forward**	**Reverse**	**Forward**	**Reverse**
Predictions using mobility only					
west	mask	0.386	0.242	0.359	0.321
south	mask	0.380	0.375	0.388	0.400
east	mask	0.378	0.446	0.406	0.477
east	no mask	0.371	0.405	0.404	0.343
middle	mask	0.386	0.385	0.426	0.500
middle	no mask	0.332	0.386	0.366	0.363
Predictions using mobility + CAMS (mean only)					
west	mask	0.402	0.312	0.397	0.451
south	mask	0.361	0.364	0.382	0.355
east	mask	0.375	0.438	0.423	0.536
east	no mask	0.381	0.378	0.418	0.417
middle	mask	0.396	0.397	0.481	0.543
middle	no mask	0.348	0.389	0.410	0.402
Predictions using CAMS (mean only)					
west	mask	0.416	0.467	0.404	0.445
south	mask	0.407	0.381	0.380	0.342
east	mask	0.416	0.523	0.409	0.503
east	no mask	0.384	0.387	0.390	0.387
middle	mask	0.454	0.526	0.444	0.544
middle	no mask	0.418	0.456	0.410	0.419

Table 2: Results split by region and mask mandate. The "Forward" prediction is the prediction of the rest of the USA by the data from this region, and the "Reverse" prediction is the opposite.

4 DISCUSSION

Given the high variability of COVID-19 case reporting trends, lockdown measures and mask mandate, even within the USA, it is not very surprising that the weather and air quality data serves as a poor indicator of COVID-19 case trends by itself. What is surprising, however, is that it provides a significant improvement over the prediction using mobility data. Indeed, it would be expected that mobility data serves as a reliable predictor of COVID-19 trends, as increased mobility is an indicator of increased interpersonal contacts, and therefore of opportunities to contract the virus. However, it can be argued that the data on weather and air quality also encodes the information on mobility, as increased pollutant emissions are indicators of increased industrial activity and vehicle traffic. Therefore, in addition to its already known effects on respiratory disease [9] [5], air quality data also encodes the information contained in mobility data, thus explaining the comparable accuracy of CAMS + mobility data and the worse performance of mobility data alone.

5 LIMITATIONS

There were many limitations to our approach that influenced the success of our models.

First of all, the COVID-19 data was very noisy. There is a large variety in criteria for reporting in different countries and even in different states. We tried to address this by focusing on the United States so the errors would be somewhat uniform. While the CAMS data is very accurate, changes in environmental data are typically a long term trend. This, along with the fact that COVID-19 is a recent disease, made it very difficult for our model find a correlation between environmental changes and COVID-19 propagation. While

Another limitation we faced while working on this project was the number of confounding factors that have an impact on the infection rate of COVID-19. Even though we sampled our data to limit the effect they might have, there's still some confounding factors we did not consider for various reasons that could have an impact over our results. Namely, the seasons, whether people respect the government recommendations, the social context (e.g. Black Lives Matter Riot, US 2020 Election), asymptomatic infections, among others could have a serious (positive or negative) impact over our results.

6 CONCLUSION

COVID-19 and environmental change are two of the biggest problems of the decade. With COVID-19 being on such a short timescale and climate change being years worth of destructive human behaviour towards the environment, it is no surprise that drawing inferences between the two is difficult. Nonetheless, we should strive to understand both of these problems as well as their interactions in order to find proper solutions for both climate change and COVID-19 pandemic. We attempted to create models such that these inferences could be made. However, the fact that CAMS data only gave a better prediction than the prediction were made by both CAMS and Mobility data is interesting to notice since CAMS data also combine within itself some Mobility data. Indeed, air quality, especially within high population density counties, is extremely affected by car mobility.

References

1. Gorry, P. A. General least-squares smoothing and differentiation by the convolution (savitzky-golay) method. 570–573.
2. Karaivanov, A., Lu, S. E., Shigeoka, H., Chen, C., and Pamplona, S. Face masks, public policies and slowing the spread of covid-19: Evidence from canada.
3. Lolli, S., Chen, Y.-C., Wang, S.-H., and Vivone, G. Impact of meteorology and air pollution on covid-19 pandemic transmission in lombardy region, northern italy.
4. Rashed, E. A., Kodera, S., Gomez-Tames, J., and Hirata, A. Influence of absolute humidity, temperature and population density on covid-19 spread and decay durations: Multi-prefecture study in japan.
5. Sharma, S., Chandra, M., and Kota, S. H. Health effects associated with pm 2.5: a systematic review. 1–23.
6. Wahltinez, O., Lee, M., Erlinger, A., Daswani, M., Yawalkar, P., Murphy, K., and Brenner, M. Covid-19 open-data: curating a fine-grained, global-scale data repository for sars-cov-2. Work in progress.
7. Wu, X., Nethery, R. C., Sabath, M. B., Braun, D., and Dominici, F. Air pollution and covid-19 mortality in the united states: Strengths and limitations of an ecological regression analysis.
8. Álvaro Briz-Redón, and Ángel Serrano-Aroca. The effect of climate on the spread of the covid-19 pandemic: A review of findings, and statistical and modelling techniques.
9. Álvaro Meca, A., Sánchez-López, A., Resino, R., Tamayo, E., and Resino, S. Environmental factors are associated with hospital admissions for sepsis-related pneumonia: A bidirectional case-crossover design. 110102.

This paper has been typeset from a TEX/LATEX file prepared by the author.