Mathematics and Computer Science
2020; 5(1): 14-30
http://www.sciencepublishinggroup.com/j/mcs
doi: 10.11648/j.mcs.20200501.13
ISSN: 2575-6036 (Print); ISSN: 2575-6028 (Online)

Towards Cryptanalysis of a Variant Prime Numbers Algorithm

Bashir Kagara Yusuf[1, *], Kamil Ahmad Bin Mahmood[2]

[1]Department of Computer, Ibrahim Badamasi Babangida University, Lapai, Nigeria

[2]Department of Computer and Information Sciences, Universiti Teknologi Petronas (UTP), Bandar Seri Iskandar, Malaysia

Email address:
bkyusf@gmail.com (B. K. Yusuf), kamilmh@utp.edu.my (K. A. B. Mahmood)

[*]Corresponding author

To cite this article:
Bashir Kagara Yusuf, Kamil Ahmad Bin Mahmood. Towards Cryptanalysis of a Variant Prime Numbers Algorithm. *Mathematics and Computer Science*. Vol. 5, No. 1, 2020, pp. 14-30. doi: 10.11648/j.mcs.20200501.13

Received: January 10, 2020; **Accepted**: January 31, 2020; **Published**: February 13, 2020

Abstract: A hallmark of prime numbers (*primes*) that both characterizes it away from other natural numbers and makes it a challenging preoccupation, is its staunch defiance to be expressed in terms of *composites* or as a *formula* listing all its sequence of elements. A classification approach, was mapped out, that fragments a prime into two: its last *digit* (*trailer* - reduced set of residue *{1, 3, 7 and 9}*) and the other *digits* (*lead*) whose value is incremented by either *1, 2 or 3* thus producing a modulo-3 arithmetic equation. The *algorithm* tracked both *Polignac's* and modified *Goldbach's* coefficients in order to explore such an open and computationally hard problem. Precisely *20,064,735,430* lower primes of *digits 2 to 12* were parsed through validity test with the powers of 10 primes of *Sloane's A006988*. Adopting at most cubic terms of predictors (as the next logical step of *Euler's* quadratic formula for primes) in multiple linear regression analysis, the generated outputs were analyzed to aid in building *Akaike Information Criterion (AIC) best model* with forward selection strategy. The main task was fragmented into atomic units of similar instances and types (an atom is a table of length *4,493,869* integer sequences where a database contains 30 relational tables with facilities for further reprocessing). A node, that supports parallel processing, stores 30 contiguous databases, and explores *4,044,482,100* successive integers. *513,649,226,700* lower natural numbers were explored by *127 hypothetical nodes* yielding primes stored in *114,300 tables* spread across *3,810 databases*. Veriton S6630G computer system with *7.86GB usable memory and processor* Intel(R) Core(TM) i7-4790 CPU @ *3.60GHz* were amongst the remarkable resources. Contrary to the apparent chaotic camouflage of primes as a bundle, the partitioned sample spaces reveal some remarkable *patterns* in terms of *intervals* of both sequence numbers and distances of separation from their immediate neighborhoods.

Keywords: AIC Best Modeling, Algorithm Analysis, Euler's Formula, Primes Pairs

1. Introduction

Excepting the even number 2 and the only special odd prime 5, it was noticed that all primes terminate with either 1, 3, 7 or 9 as its last digit. It is this pattern the study intends to address [27]. Manipulation techniques that involve primes are certified to provide high degree of security since the likelihood of finding a back door has been spread to an extremely large search space with a lone solution that hardly occurs by chance [4, 18]. Besides, a fair insight into the behavior of primes pertaining to: distance of separation between prime pairs in the context of Polignac and Goldbach coefficients is to be sought after, Lack of predictable analytical tool makes primes so complex that an in-depth investigation is necessary to expose their behavioral laws. Primes are used in concealment of sensitive data and messages in open and insecure environment (Figure 1 visualizes the scenario).

Figure 1. *Picture of decent dressing versus nudity that are analogous to cipher-text and plain text in cryptography, respectively.*

The basic primality testing algorithm that had withstood the test of time in terms of reliability, data integrity and precision is the *sieve* of Eratosthenes which confines the search for composites of n to the range: $2 \leq p \leq \sqrt{n}$. An algorithm was developed that considers both reduced set of residues (i.e., *Trailers* = {1, 3, 7, 9} depending on value of the last *digit* of *n*) and some arithmetic operation on the other *digit*s modulo 3 (called *Cycle*, being equal to zero). Although knowledge of reduced set of residue has long been established, attention of researchers was not focused on its interactions with the rest of other digits of primes. Past research efforts were devoted on primal value as a whole without venturing into fragmentation as embarked upon by this work. Extant literature is silent on this aspect. This research seeks to identify the impact of digits fragmentation and distance of separation (by both value and sequence numbers) on the behavior of primes. A demonstration of how both the new variables of prime numbers and existing cryptographic terms inter-relate is given by Figure 2 below:

Figure 2. *Pictorial representations of the newly introduced prime variables introduced and their interactions with existing terms of cryptography.*

Reference [14] alerted that with security and privacy of digital money and commercial applications totally entrusted in the care of encryption governed by prime numbers, any major breakthrough that predicts the orderly sequence away from a game of chance will bring catastrophic consequences to the financial world [5–6, 12]. Using sufficiently large prime numbers (e.g. 100 *digit*s) as private and public keys, the task of factoring semi-primes $n = p \times q$ for some prime *p* and prime *q* (e.g., in *RSA* cryptosystem), is infeasible to undertake given all the computational power available [17]. The work is made so expensive that it is unsolvable in polynomial time ($n!$ *or* k^n *where* $k > 1,\ n \geq 10$ search space as an adversary) [19].

Euler's quadratic function generates only prime numbers for n=0:39 and is a polynomial in terms of only one variable (var). As the next logical step, this study will explore the cubic function in searching for the existence of that function that not only generates primes but also never yields a composite. Also, this work is not restricted to just one variable, it expands the domain to multivariate polynomials up to a limit of six independent variables each of a maximum of third degree of terms. Moreover, proving conjectures is a research area that is open for contribution with some of these problems having stagnated over many decades [20, 27]. Is it possible to discover a polynomial which computes all possible prime numbers, exhaustively without skipping of some sequences, in error? Euler's function is similar to 6th cyclotomic polynomial whose constant term is not unity [6, 7]. The 6^{th} cyclotomic polynomial is an irreducible and unique polynomial with coefficient terms as integers that divides $x^6 - 1$ and can never divide $x^k - 1$ *where* $k = 0:5$.

The sole goal of the research effort on the subject matter (primes) aims to address the main problem "all polynomials generate composites - they are unreliable since they always inevitably lead to some errors of prime computation" [11, 19]. Research questions that help navigate this study, strive to verify newly designed and variant algorithm of primality test against standards with a view to seeking answers to the following:

What criteria aids to simulate an approximately "hard", "easy" or "average" test of primes' discovery?

Which parameters of efficiency analyze and evaluate this algorithm against its pairs for testing primality?

Is it possible to build a more robust and fault-tolerant model in terms of errors of estimation?

The paper's central research objectives are:

To identify some basis for proving conjectures and persistently hard and open problems as regards primality testing.

To search for such a polynomial function that formulates all possible prime numbers reliably with tolerable errors?

To explore existence of that polynomial of degree more than the quadratic function for the existence of a prime formula.

The rest of the paper has been structured into 6 sections. First, a review of the existing and related literature most suited to the prime numbers' study was carried out in section 2. Next (i.e., section 3), materials and design phase was discussed with a view to outlining the conceptual framework and direction of the research. Then in section 4, normality test conducted

justifying why this quantitative research has passed parametric statistical analysis conditions. As part of the research methods and procedures used, AIC model building formed the next section with exhaustive discussion of multiple linear regression versus Akaike's analyses (as in section 5). This is followed by the results and its discussion (in section 6), showing relevant findings, analysis and summary. Finally in section 7, this article concludes by stating the implications, limitations and charting new directions for further research.

2. Related Work

The traditional approach of testing primality is by "trial division". Observations reveal that the last *digit*s of all primes are odd numbers with the exception of 5 whilst 2 is the only even *prime*. Factorization strategies also showed that the scope of work is further downscaled to only those odd numbers as large as $\sqrt{n}$ since this happens to be the largest composite of an integer *n* if at all such a factor exists. The complexity of this algorithm, *log (n)*, is not described by a polynomial function. This forms the foundation of reliability of generation of prime numbers. However, with the aim of improving efficiency this basis has attracted a lot of improvised techniques of primality testing. Efficiency experts found that this testing technique is grossly inefficient and there exists a vacuum for further development [27]. Revolutionary achievement was realized by Pierre de Fermat who stated that for any *prime* integer *n* and *a* both of which have no common divisors, Equation (1) suffices thus:

$$a^{(n+1)} \equiv 1 (\mathrm{mod}\, n) \quad or \quad a^n \equiv a (\mathrm{mod}\, n) \tag{1}$$

A class of computational problems, called *NP*, decides (Yes-No) response for which a Yes-certification is needed and checked in time that is a polynomial in size of the input. In the mid-1970s, [22] conjectured that primes are in NP computational class of problems by showing certification through repeated Lucas-Lehmer testing and this was achieved using no more than about *4logN* bits and checks were performed in no more than about $\log^3 N$ steps. Obstacles in the problem classification exists especially the difficulty of finding suitable primitive root and factoring of *n-1*.

Reference [7] attributed to Eduardo Lucas the proof of the following theorem.

Theorem 1. If both '*n'* and '*a'* are integers, where $n > 1$, and $a^{n-1} \equiv 1 \pmod{n}$ but $a^{(n-1)/q} \not\equiv 1 \pmod{n}$ for each *prime q* not a factor of (*n—1*), *n* is then *prime*.

Lucas—Lehmer primality test was a hybrid adaptation of this proved theorem by [23]. Lucas-Lehmer's test is not effective as general-purpose primality tests because of errors discovered. However, according to [7, 6] combined primality tests of both Lucas-Lehmer and cyclotomic one to design a hybrid test that proves very effective.

Let $\pi(x)$ be the count of primes $p \leq x$, and let $Li(x) = \int_{t=0}^{x} \frac{dt}{\ln t}$. Then, $\pi(x) = Li(x) + O\left(\sqrt{x}\log(x)\right)$.

Next, [29] developed a randomized primes algorithm that has a probability of error with minimal optimization requirement for all inputs, proving primes are in bounded – error probabilistic polynomial time (BPP). Moreover, the probability of the error in the opposite direction is arbitrarily small renders it not very reliable. However, the error occurs always in the opposite direction (composites) but never in primes. Also, [29] states that:

Theorem 2: Suppose that *n* is odd, and also assume that '*a*' is a positive whole number $\ni \; a \not\equiv 0 \pmod{n}$. Then, $a^{(n-1)/2} \equiv \left(\frac{a}{n}\right) \equiv \pm 1 \mod n$, but $\frac{a}{n}$ represents the Legendre symbol defined as either 1 if '*a*' is a quadratic residue modulo *n*, or *0* if $n \mid a$, or otherwise it is -1.

Rabin, as re-published in [17], later modified [26]'s algorithm presenting it as a randomized and unconditional algorithm of polynomial time order. He also proved that primes belong to *co-RP* computational class of problems. It is widely used in practice and called Miller-Rabin`s primality test..

Besides, [2] proposed a deterministic primes algorithm (called cyclotomic method) which executes in $k.\log\left(n^{c\log^3 n}\right)$ time bound, for the constant pair *c* and k. Later, Cohen and Lenstra reduced the algorithm into its simple form. Its efficiency of execution is very much noticeable in integers of several hundreds of *digit*s long especially on fast workstations. The trio postulated thus:

Theorem 3: There are positive, absolute, calculable constants c_3 and $c_4 \ni$ for each $n > 100$, $\log\left(n^{c_3\log^3 n}\right) < g(n) \leq f(n) < \log\left(n^{c_4\log^3 n}\right)$.

Their algorithm has been strongly challenged by criticism of experts that it is both prone to bugs that are very hard to detect and its program implementation though possible but proves very difficult to actualize.

Most recent development in primality testing was a variant of [3] that achieves a truly polynomial bound time of $(\log n)^6.(2+\log^2 n)^c$, c is a real number and is calculable in practice effectively [8]. Elsewhere [7] has argued that *AKS* algorithm is yet to accomplish an effective test?

Comparatively, *ECPP* algorithm is faster than [3] with the latter being not a practical algorithm. Relatively, [15, 18] is better in terms of speed of execution though with attendant cost of minimal and probable error.

Moreover, recently [30] achieved a breakthrough in formalizing the proof of the twin *prime* conjecture thus:

$$\lim_{n\to\infty} \inf(p_{n+1} - p_n) < 7 \; x \; 10^7 \tag{2}$$

It is then clear that an in-depth investigation becomes a necessity in this search for more knowledge about the behavioral patterns of primes.

3. Materials and Design Phase

Some interesting pattern associated with decomposition into last and perhaps the most significant digit (called trailer) and other digits (referred to as lead) is captured by the (3) as function given below:

$$P = \begin{cases} S_n = \begin{cases} a_n = 2, & if \quad n = 1 \\ a_n = a_{n-1} + n - 1 & n = 2,3 \\ a_n = 7^{n-3} & n = 4 \end{cases} \\ T_n = b_n.c_m = \begin{cases} b_n = 3(n-1), & n \geq 1 \\ . \\ c_m = 7^{m-1}, m = 1,2; & (b_m \% 7 \neq 0) \end{cases} \\ U_m = d_n.e_m = \begin{cases} d_n = 3n - 2, & n \geq 1 \\ . \\ e_m = \begin{cases} 7^{m-3}, & m = 4 \\ . \\ 3^{m-1}, & m = 1:3 \end{cases} \\ . \\ e_m = \begin{cases} 7^{m-2}, & m = 3 \\ . \\ 3^m, & m = 1,2 \end{cases} \\ . \\ e_m = 3^{m-1}, & m = 1:3 \end{cases} \\ V_n = f_n.g_m \begin{cases} f_n = 3n - 1, & if \; n \geq 1 \\ g_m = 3^m & m = 1,2 \end{cases} \end{cases} \tag{3}$$

$$S_n = (s_1, s_2, s_3, s_4) = (2,3,5,7) \quad 1\,digit$$

T_n, U_n, *and* V_n all consist of more than 1 *digit* terms except trivial case of 07 in $T_1 = b_1.c_1$. Otherwise all other terms are more than single *digit*s e.g., 11, 13, ⋯. c_m, e_m and g_m are the *Trailer*s while b_n, d_n and f_n represent respective leads and *Trailer* = {1, 3, 7, 9}. Coupled with analysis of data obtained from variants of Euler's formula and arithmetic progression properties, an initial algorithm needs to be developed from (3) with derivatives from Perl syntaxes and necessary variables. The recurrence relation in (3) transforms into the data dictionary as shown in Table 1. (Note: Subscript n concerns indices for *Trailers* whilst m refers to *Cycle* variable and its attendant indices.)

Table 1. *Description of the variables considered for features extraction.*

Variable	Description
SNo (X_1 /SNo)	Initial identification number assigned to the observed population data by the Algorithm
Digits (X_2 /Dgt)	Number of digits found in a given prime number. e.g., for the Prime = 111, Digits = 3 whilst Prime = 7 has Digits = 1.
Trailer (X_3 /Trl)	Last digit in a prime number. E.g., for Prime = 23, Trailer =3; Prime = 17, Trailer = 7; Prime = 119, Trailer = 9 etc.
Trailer2 (X_4 /Tr2)	Last 2 digits in a prime number. E.g., for Prime = 331, Trailer2 =31; Prime = 1051, Trailer2 = 51; Prime = 19, Trailer2 = 19 etc.
Tr2_Sta (X_5 /Tr2)	Tr2_Sta=1 when Trailer2 ϵ P, otherwise (i.e., Trailer2 ϵ C) Tr2_Sta = 0. E.g., For Prime=739, Tr2_Sta=0 because Trailer2=39 ϵ C.
Lead (X_6 /Lid)	All digits of prime except the Trailer, e.g., if Prime = 571, then Lead = 57.
Cycle (X_8 /Cyc)	If (Lead + j) mod 3 = 0, then Cycle=j, j=1:3
Polignac (X_9 /Pol)	The positive difference between this prime and its immediate predecessor. e.g., for Prime = ans = 53, prev = 47, Pol = (ans — prev)/2 = (53 — 47)/2 = 3.
Pol_Grp (X_{10} /PGp)	If (Polignac = 1) & (Cycle = k), then Pol_Grp = k, k=2,3; Otherwise (Polignac >1) & (Cycle = k) Pol_Grp = j, j=3+k.
Goldbach (X_{11} /Gol)	The difference between this prime and its three immediate predecessor. e.g., for Prime = ans = 53, prev=47, prv2=43, prv3=41, Gol = (prev + prv2) — (ans + prv3) = (47+43) — (53+41) = 90 — 94 = —4;
Gol_Grp (X_{12} /GGp)	If (Goldbach = 0) & ((Polignac + k)% 3 = 0), then Gol_Grp = k, k=1,3; otherwise (Goldbach <> 0) & ((Polignac + k)% 3 = 0), Gol_Grp = j, j=3+k.
Packets (X_{13} /Pkt)	A bundle (Packet) of distinct logical operation as executed by the algorithm during a given prime number generation.
Sieve (X_{15} /Siv)	Count of Prime numbers within the sieve of Eratosthenes, i.e., for any n, relative primes from 2 to sqrt (n)
Discarded (X_{16} /Dsc)	The positive difference between sqrt (n) and sieve. e.g., for Prime = ans = 47, Dsc=round(sqrt(47),0) —Siv=7—4=3;
L2G_ZZZ (X_J /ZZZ)	L2G_Bm2 = Lead2 - prev_ZZZ, where ZZZ = TrJ (Trailer=J) [J = 1, 3, 7, 9]; T2C (Tr2_Sta=0), T2P (Tr2_Sta=1); CyJ (Cycle=J) [J = 1:3]; TCJ (Polignac=1 & Cycle=J) [J = 2:3], Pm?(Polignac>1 & Cycle=J) [J=1:3]; GmJ (Goldbach=0 & (Polignac + J) mod 3 = 0), BmJ (Goldbach<>0 & (Polignac + J) mod 3 = 0) [J = 1:3]; T?C(Tr2_Sta=0, Trailer = J), TJP(Tr2_Sta = 1, Trailer = J) [J = 1,3,7,9]; ALL(no condition imposed); [16 < J < 46]

A general polynomial class that resembles Euler's function is the 3rd (i.e., n=3 is *prime* $\phi_3(x)=\sum_{i=0}^{3-1} x^i = x^2+x+1$) and 6th (*n=2p*, where *p=3* is also *prime* $\phi_{2(3)}(x)=\sum_{i=0}^{3\_1}(\ \ x)^i = x^2\ \ \ x+1$), cyclotomic polynomial with the constant term not unity but some other integers as enumerated in (4). It is interesting to observe that there is no cyclotomic polynomial of degree 3 (the cubic function strictly) [6, 7]. The variants of Euler's function, listed in (4), were obtained after some trial and error process [11]. Euler's variants discovered are given by (4) below:

$$y = x^2 \pm x + E, \quad . \quad .. \quad . \tag{4}$$

Where E = {A, B, C, D} and
A = {11, 41, 71, 101, 131},
B = {19, 49, 79, 109, 139, 169, 199, 229, 259, 299},
C = {17, 107, 137, 167} and D = {-13, -43, 173, -103}.

Moreover, variants of the linear form were explored and are related by (5) thus:

$$y = bx + a \quad where \quad a = \{P, Q, R\} \tag{5}$$

b = {10, 26.25, 30, 35, 210},
P = {-122, -45, -41, -38, -10, -4, 1, 11, 53},
Q = {-127, -79, -43, -41, -23, -11, -1, 11, 53},
R = {-77, -53, -19, -17, -7, 1, 11, 19, 79}.

In addition to being arbitrarily distributed, [9] also observed that prime numbers behave like weeds that apparently disobey all natural rules except those of random chance and are chaotically distributed, flaunting all discernible order for their generation. Let us consider the odd numbers 1 - 100 below:

P = {03, 07, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89, 97}

C = {09, 15, 21, 25, 27, 33, 35, 39, 45, 49, 51, 55, 57, 63, 65, 69, 75, 77, 81, 85, 97, 91, 93, 95, 99}

Surely, these randomization conflicts of including both composites and primes in the last 2 odd digits is a mystery worth investigating - hence the exploration of the parameter Trailer2 Status (Table 2 refers) with the aim of discovering if any pattern can be established in the sample space so formed [2].

Among the list of parameters discovered during algorithm design, Goldbach and Polignac constitute the main factors for classification of population data were split into the sample spaces as defined by (6) and (7) thus:

$$Polignac = (\Pr imes - prev) / 2; \tag{6}$$

$$Goldbach = ((prev + prv2) - (Prmes + prv3))/2 \tag{7}$$

Where *prev, prv2* and *prv3* are successive previous prime numbers already processed in that order.

Extensive coverage by the literature reveals the following parameters of interest applicable to the description of this algorithm for prime numbers generation as shown by Table 1 below whilst Figure 3 gives the pseudo code of the algorithm:

```
Create List for Sieves in a Hash Data Structure (@p_list) - Primes Maximum of 1,182,869
Set Up Parameters for Basis Sn = {2, 3, 5, 7} And Transition to Next Machine – For Smooth Take Off
For ($i=$initial;$i<=$final-1;$i++)     # Partitions To Enforce Parallelism Within A Machine
        For ($nach=$start; $nach<$stop; $nach=$nach+2)     # Progressing To Search Next Odd Integer
                If ($nach > 9 && $nach < 1399161628201)     # Limit of Integers to Be Tested for Primality
                        $n = ($lead + $j) % 3;       # A Modulo 3 (Compound Statement) Arithmetic with Remainder = 0
                        # $k = 0; $k = 1 if (($nach ± 1) % 6 == 0); # All Primes of Digits > 1 Must Satisfy This Condition
                        If ($n == 0 && $k == 1)     # Modulo (3 And 6) Arithmetic Filters Potential Primes Only
                                Do Case # Choosing One of Three Options But For Loop, Rotating $J, Is Needed To Implement
                                        Case ($j == 3)     # $lead = {3, 6, 9, 12 ...}
                                                If ($trailer == 1 || $trailer == 7) # Tn = bn, cm in (3)
                                                        If (($lead % 7) != 0 && ($trailer != 7)) # $lead = {3, 6, 9, 12, 15, . . . }
                                                                # Excludes $Lead Divisible By 7 {217, 427, 637, 847, 1057 ...}
                                                                House_Keeper ();
                                                # Else $trailer = {3, 9} (digits sum) % 3 = 0 e.g., 33, 39, 63, 69, 93, 99 …
                                        Case ($j == 2)     # $lead = {1, 4, 7, 10 …} and $trailer = {1, 3, 7, 9}; Un = dn, em in (3)
                                                House_Keeper ()
                                        Case ($j == 1)     # $lead = {2, 5, 8, 11 ...}; $trailer = {1, 7} (digits sum) % 3 = 0
                                                If ($trailer == 3 || $trailer == 9) # Vn = fn, gm in (4)
                                                        House_Keeper ()
                Reset Appropriate Variables to Advance to another Output File
Declare Success and Prepare For the Next Category of Records
Procedure House_Keeper {
        Repeat Trial Divisions (relative primes within a given Sieve, intermediate (Residuals, Quotients)) Until (Satisfied)
        If a factor is not found within a Sieve of Eratosthenes then $nach is Confirmed to be Prime else it is a Composite
        Adjust and Save Variables Appropriately
}
```

Figure 3. *Primes generation algorithm – trial division variant implementing the recurrence in (3) above (Perl-like syntax and pseudo code).*

Also, L2G_TC2 measures respective distance between these successive twin primes is tracked by (8). Similar computation for Goldbach gaps can be traced to Table 1 above.

$$Pol\_Grp = 2 \quad if \quad (Polignac = 1) \ \& \ (Cycle = 2)$$
$$Pol\_Grp = 3 \quad if \quad (Polignac = 1) \ \& \ (Cycle = 3) \ . \quad (8)$$

Errors of primality tests occurred in both the algorithm under development and the new variants of Euler's formula being discovered but with persistent verification strategies, especially the conformity with the 10^Nth primes (*for N=0:10*) of *Sloane's A006988* [18] (Table 2 suffices), the computation was error free. In addition, Table 3 shows further extension of the 10^nth primes (*for N=11:15*).

Table 2. *Verification of program by showing conformity to the Sloane's A006988 (10^N) th prime for N=1:10.*

N	1	2	3	4	5	6	7	8	9	10
(10^N)th prime	29	541	7,919	104,729	1,299,709	15,485,863	179,424,673	2,038,074,743	22,801,763,489	252,097,800,623
Mean gap	3	6	8	11	13	16	18	21	23	26

Table 3. *Extension of verification of Sloane's A006988 (10^N) th prime for N=11:15 for future verification.*

N	11	12	13	14	15
(10^N)th prime	2,760,727,302,517	29,996,224,275,833	323,780,508,946,331	3,475,385,758,524,520	37,124,508,045,065,400
Mean gap	28	30	33	35	38

In this quantitative research, numeric values that can undergo arithmetic operations like computing averages, measures of dispersion of the data (e.g., variance, standard deviation etc.) are mainly considered [25]. The population data space is segmented into groups of both lower (L) and upper (U) segments with 500,000 records each as analyzed for sampling purposes (see Table 4).

Table 4. *Distribution of samples versus population data spaces earlier earmarked.*

Digits (Dgt)	Dgt2-7	Dgt08	Dgt09	Dgt10	Dgt11	[a]Dgt12[a]
% of population	100.00%	19.62%	2.22%	0.25%	0.03%	0.01%

[a] Dgt12L=First 500,000 records of Dgt=12, U=Last 500,000 records; similarly for others

This research work has its main purpose as to establish relationship between the different predictor variables and the *prime* parameter. Regression analysis requires data collection on the independent variable whose values are assumed to influence a dependent (response) variable. A linear form is assumed for cause and effect relationship [10].

In order to solve complex problem of exploring about *half a trillion* integer sequences in the program design, a problem solving strategy that breaks a giant main problem into smaller and more manageable units of similar instances of the main type was implemented until a sub-problem is found that can no longer be subdivided (becoming *atomic*) and a solution of the atomic type is found out rightly. By combining together instances of partial answers, the overall solution is formed. The units of processing is output files (text) each of length *4,493,869* sequences of integers searched for primes and transferred as tables in *MS Access* databases (db). An Access database consists of *30 tables* with provision for queries and other database objects to support further reprocessing. A hypothetical computer, called a node, in a super computer that supports parallel processing houses *30* databases equaling *900* text files each in a separate table. In essence, for a given node in the cluster, *4,044,482,100* successive integers are searched for primes. 127 such nodes translate to a total of *513,649,226,700 integer sequences* producing exactly *20,064,735,430* primes housed in *114,300* tables resident in *3810* databases.

Due to storage astronomically high demand for both storage and processing requirements, Hewlett Packard (*HP*) -server could not handle this computation in the midst of contending users and other applications. As such, stand-alone computers were deployed to support the parallel processing at a great cost to human coordination of tasks and computing resources. In addition, program extraction and reporting language (*Perl*) was chosen to implement the algorithm with some aid from *VB script* to help interface *Access and Excel* applications. The available computing resources are *Veriton S6630G* computer system supporting *64-bit operating system with 8.00 GB* (only *7.86GB* usable) memory and processor Intel(R) *Core(TM) i7-4790 CPU @ 3.60GHz*. Parallelism and distributed processing of the algorithm implemented on standalone computers at the 3 computer Labs as shown in Figure 4.

Figure 4. *A typical computer center depicting the coordination and integration of 127 standalone systems for parallel processing.*

Furthermore, this algorithm's performance needs to be further compared with traditional *sieve* of Eratosthenes as can

be seen in Table 5. Also, distance separating successive *prime* pairs, in terms of both *Polignac* and *Goldbach* constant coefficients at their extreme boundaries, across all the *digit*s has been epitomized by Figure 5. Observe the divergent nature of the growth of the plots with Goldbach covering both positive and negative spectra but Polignac remaining on the more realistic positive spectrum (Figure 5 portrays the relationships). This suggests a strong relationship (link) between distances of separation (directly proportional) to number of digits (running into thousands) which in turn explains Zhang's wide choice of interval of *70000000* between twin primes.

Table 5. *Analysis of population data space by digits, gain & loss in trial divisions and cost of attendant logical comparisons.*

[a]Dgt	Actual Trials([a]tdiv)	Avoided Trials ([a]gain)	Mean tdiv	Mean gain	% benefit	Total Comparisons ([a]tcmp)	Mean comparisons	[a]cmp per tdiv
2	75	78	5	6	54.55%	345	5	21
3	1,186	2,044	12	21	63.64%	3,694	4	143
4	20,731	54,576	28	73	72.28%	51,278	3	1,061
5	408,188	1,466,193	70	248	77.99%	893,731	3	8,363
6	8,651,274	40,211,580	178	823	82.22%	17,939,951	3	68,906
7	195,673,122	1,120,418,401	471	2,693	85.11%	396,767,502	3	586,081
8	4,619,337,276	31,620,653,730	1,275	8,726	87.25%	9,285,820,658	3	5,096,876
9	113,012,158,128	901,753,432,230	3,522	28,102	88.86%	226,441,363,353	3	45,086,079
10	2,844,559,625,930	25,947,816,765,154	9,880	90,121	90.12%	10,107,543,261,041	4	404,204,977
11	73,232,673,298,300	752,425,838,916,472	28,049	288,180	91.13%	408,864,210,532,525	6	3,663,002,301
12	636,385,190,033,576	8,009,519,891,461,860	52,753	663,942	92.64%	2,046,792,337,342,650	4	15,946,680,618

*[a]Dgt=Digit, tdiv=trial divisions actually executed, gain= trials avoided, mean tdiv = (tdiv/(tdiv+gain))*sqrt(highest prm), mean gain = (gain/(tdiv+gain))*sqrt (highest prm), benefit = (mean tdiv/(mean tdiv+mean gain))

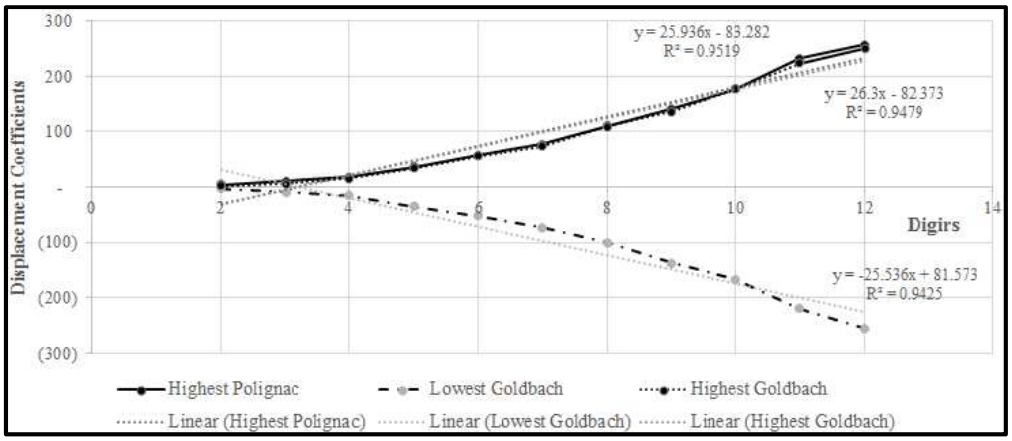

Figure 5. *Growth of primes based on distance of separation from its neighbors (Polignac and Goldbach coefficients) against digits.*

4. Normality Test

Parametric methods are more assumptions concentric and when they are correctly made, the methods produce more precise estimates. On the contrary, they are also very unreliable and misleading when the assumptions are not right. The normal family of distributions all have the same shape (bell-like) and are characterized by mean, standard deviation and variance - meaning that knowledge of these measures in a normally distributed probability space, estimating future values is guaranteed as robust with high degree of accuracy and security [10].

Polignac conjecture involves successive primes and is expressed by the recurrence relation as shown by (9) [20]:

$$a_n = a_{n-1} + 2k, \quad where \quad k > 0 \tag{9}$$

On the other hand, critical appraisal reveals that *Goldbach*'s conjecture on *gap*s among three primes can be simplified to follow the recurrence relations (10) below:

$$a_n = \begin{cases} 1 \ \ if \ \ n = 1 \\ . \\ 3 \ \ if \ \ n = 2 \\ . \\ 5 \ \ if \ \ n = 3 \\ . \\ a_{n-1} + a_{n-2} - a_{n-3} \pm 2k \ \ where \ \ k > 0, and \ n > 3 \end{cases} \tag{10}$$

These *gap*s are subjected to measures of statistical spread so as to observe whether the observations on primes are normally distributed or not before parametric testing is done [27].

An informal test of normality on generated sample data was done by the use of histogram against normal probability curve, as shown by Figure 6. With sample data running into 20 billion of observations and the need to capture the structure of a model in a fixed bell-like shape for the normal distribution [25], use of measures of dispersion is resorted to in order to dissect through all classes of the sample data as expressed in

*gap*s of both *Polignac* and *Goldbach*'s coefficients. The symmetric query (i.e., both *Polignac* and *Goldbach* are interchangeable variables) below confirms that the data set is normally distributed and its data is portrayed by Figure 6.

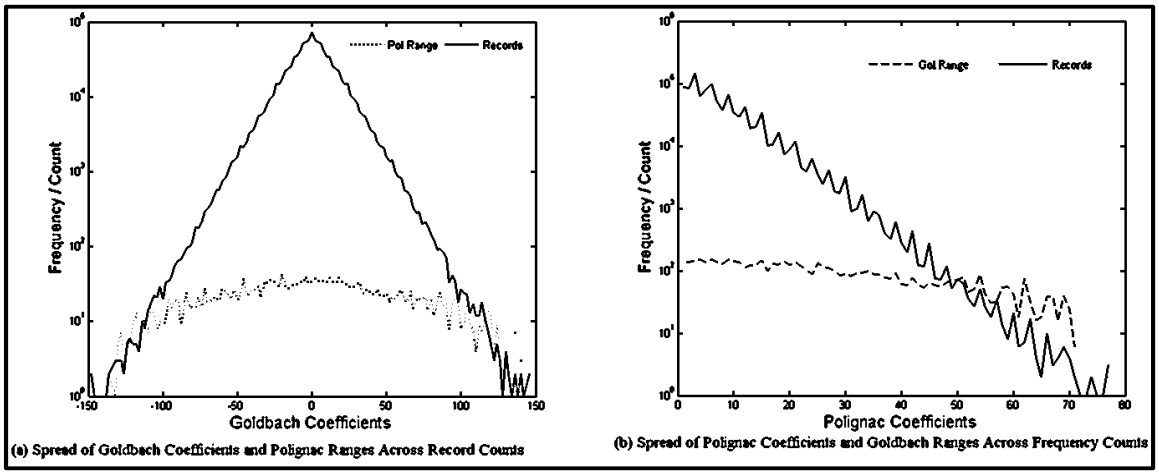

Figure 6. *Growth of prime numbers based on distance of separation from its neighbors via Polignac and Goldbach coefficients.*

5. AIC Model Building

Regression of the simplest form is represented by a linear equation in which a response Y is measured, for every instance of observations, by some relation to a predictor (regressor or covariate) X [16]. However, there are cases where by more than one predictor (Xi variables) are involved in the determination of the Y response. According to [25], observation reveal that the relationship is not perfect and as such the responses fluctuate around some mean values; yielding a model of the form of (11) - (12):

$$Y = \alpha + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_P X_{i,p} + \xi_i \tag{11}$$

$$Y_i = E(Y \mid X_i) + \xi_i \tag{12}$$

It is assumed that for the errors ξ_i, denoted by $E(\xi_i) = 0$ and $\text{var}(\xi_i) = \sigma^2$, i.e., equality of all variances [25]. The $\xi_i s$ should be mutually independent of one another and have a fairly normal distribution in a moderately small sample sizes.

There are some complexities in measurement of the coefficients, Sums of Squares of Estimation (*SSE*), Residual (*SSR*) etc. The system of linear equations is best transformed into matrix notation for ease of manipulation. In matrix form, the linear system of equations for the n observations can be reduced, very compactly, to the form (13):

$$Y_i = \beta X + \xi_i \tag{13}$$

Where Xβ denotes the product of the matrix X and vector β. On expansion of (13), it results to (14):

$$\begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{1,1} & X_{1,2} & ... & X_{1,p} \\ 1 & X_{2,1} & X_{2,2} & ... & X_{2,p} \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & X_{n,1} & X_{n,2} & ... & X_{n,p} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \\ \\ \beta_p \end{pmatrix} + \begin{pmatrix} \xi_1 \\ \xi_2 \\ . \\ . \\ \xi_n \end{pmatrix} \tag{14}$$

Where $X_{m,n}$ denotes the nth predictor variable that was measured for the m[th] observation whilst $\beta_i (i = 1:n)$ are the coefficients which are unknown and their values must be determined using least squares method.

The least squares statistical method aids us to estimate β (a *p+1*-dimensional column vector referred to as the slope-vector that includes both the intercept and the slopes). Also, striving to minimize the length of the error vector, yields (15):

$$\sum_{i=1}^{n} \left(Y_i - \alpha - \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_P X_{i,p} - \xi_i\right)^2 \tag{15}$$

An underlying assumption in this multiple regression analysis is that the predictors are linearly independent - excluding correlation among all pairs of the variables since interaction effects of a factor is quantifiably regarded as insignificant [16]. The main objective is the minimization of the sum of squared errors (16) by setting.

$$\begin{aligned} Y &= X\beta \\ X^T Y &= X^T X\beta \quad (\textit{Matrix Multiplication}) \end{aligned} \tag{16}$$

Also, by linear independence, the inverse of the multiplication of the transpose of matrix *X* and *X* itself exists. Therefore, (17) - (19) evolves:

$$\hat{\beta} = \left(X^T X\right)^{-1}\left(X^T Y\right) \tag{17}$$

Fitting the data values into (18)

$$\hat{Y} = X\hat{\beta} = X\left(X^T X\right)^{-1}\left(X^T Y\right) \tag{18}$$

While residuals are given as (19)

$$\hat{r} = Y - \hat{Y} = Y - X\left(X^T X\right)^{-1}\left(X^T Y\right) = \left(1 - X\left(X^T X\right)^{-1} X^T\right)Y \tag{19}$$

Both fitted values and residuals have no correlation since (20) is defined thus:

$$\sum \left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \overline{Y}\right) = 0 \tag{20}$$

Hence, the decomposition (21) below holds

$$SSTO = SSE + SSR \quad (21)$$

Since $SSTO = SSE + SSR$, the decomposition = sum of squares (SS) for observations suffices. Equation (21) is now equivalent to (22)

$$\sum \left(Y_i - \bar{Y}\right)^2 = \sum \left(Y_i - \hat{Y}\right)^2 + \sum \left(\hat{Y}_i - \bar{Y}\right)^2 \quad (22)$$

Table 6 presents the degrees of freedom (*DF*) together with the sums of squares above.

Table 6. *Sum of squares associations with DF.*

Source	Formula	DF
SSTO	$\Sigma(Y_i - \bar{Y})^2$	n − 1
SSE	$\Sigma(Y_i - Y)^2$	n − p − 1
SSR	$\Sigma(Y - \bar{Y})^2$	p

Further, defining mean squares as a sum of corresponding squares divided by the respective degrees of freedom takes the form:

MSTO = SSTO/(n−1); MSE = SSE/(n−p–1); MSR = SSR/p

F is also defined as *MSR/MSE* whilst the proportion of explained variation (*PVE*) is *SSR/SSTO*. *PVE* always lies in the range of *0* and *1* - values close to *1* shows a closer fit [16].

The analysis addresses uncertainty in a mathematical function (or curve) that fits data with random errors. A line equation exactly fits through any two points. Least squares method evaluates deviations from approximate solution [10].

Model building (variable selection) is a process that identifies and evaluates those relevant variables for simulation process [5]. Best fit or trend line is a straight line on a graph of scatter plot that best represents the general direction that a group of data points appear to be heading. Care must be taken in cases where the values sought after do not fall within the plotted values (extrapolation) as it is less reliable for prediction [9].

Variable selection penalizes those models of several variables that fail to better fit than fewer variable models. The Akaike Information Criterion (*AIC*) is one approach that perfectly implements this penalty favoring models with fewer variables over full models provided they respect *PVE* descending order trend [16]. *AIC* is evaluated (23) thus:

$$AIC = n\log(SSE/n) + 2(p+1) \quad (23)$$

6. Results and Discussions

6.1. Preliminaries

Extraction of sample data for the condition *Trailer*=1 is facilitated by running SQL query on a table comprising first one million primes with all relevant fields defining the object *prime* as depicted by Table 1 and extending the individual relevant predictor variables into appropriate power terms (initially set as cubic terms). The main effect terms were preferred against attendant interactions between possible pairs of variables. In most cases, more than one powers of an independent variable combine to determine the quantity of the response variable *Y*. This leads to the use of multiple regression in order to fit a model where *X* exists in more than one power transform [16].

By observation, the predictor variable '*Trailer*' has four distinct values i.e., 1, 3, 7 and 9. Table 7 is the result for the 1-variable. Let us consider the case of sample where (*Trailer* =1). For a 1-variable subset, "Gap" constitutes the best fit predictor with its *lowest AIC* value of 2.7182195E+06. Iterative computation up to 6 variables gives Table 7.

Table 7. *AIC results for a typical one-variable subsets.*

PREDI-CTORS	F	PVE	AIC
Pkt	2.6745765E-09	2.2276933E-04	3.3268106E+06
Trl	3.7328262E-06	2.3721287E-01	3.2974434E+06
Pol	2.3927317E-03	9.9500845E-01	2.7515457E+06
Gol	3.2569371E-03	9.9632806E-01	2.7182195E+06
Siv	2.6745765E-09	2.2276933E-04	3.3268106E+06
Dsc	5.4976254E-11	4.5800508E-06	3.3268343E+06

From the underlying literature, let us consider the strategy of stratifying the population data into segments of these parameters of immense interest: *Digit*s, *Cycle* and composite / prime *Trailer2*. For convenience and ease of manipulations the strata are grouped and organized thus: *Digit*s (2-7, 8, 9, 10, 11 and 12), *Cycle* (1, 2 and 3) and composite / prime *Trailer2*. For us to enclose and capture behavior at extreme conditions, the strata are further split into both lower (*L*) and upper (*U*) boundary conditions most especially for *digit*s sample space. Best *AIC* values for the *Dgt12U* sample space as classified according to subsets is depicted by Table 8.

Table 8. *AIC best models by number of subsets ranking.*

Predictors	AIC	Significant F	Rank
Tr2	3,555,871.360	0.93698204	6
Cyc, Pkt	3,603,333.986	0.77211182	11
Tr2, Cyc, Pkt	3,732,311.101	0.96750115	40
Cy2, Cyc, Tr2, Pkt	3,800,152.023	0.98895865	55
Cy2, Cyc, Tr2, Pkt, Gol	3,957,056.171	0.98872520	62
Cy2, Cyc, Tr2, Pkt, Gol, Pol	4,083,028.758	0.99430893	63

The equation derivable from Table 6 is given by (24):

$$\begin{aligned} \Pr imes = {} & 5.140169\ E11 + 2.336314\ E3Pol - 8.976387\ E1Pol^2 - \\ & 6.10939\ E-1Pol^3 - 1.770268\ E-7Pkt^3 - 8.532908E1\ \ Gol - \\ & 1.226385E0\ \ Gol^2 - 8.715591E\ \ -2Gol^3 \end{aligned} \quad (24)$$

Applying to all subsets whilst seeking for the most minimal *AIC* value representing the best model, Table 9 captured the results leading to formulae of all the sample space earmarked

Table 9. *All possible subsets for trailer sample space.*

	Cardina lity	Subsets
0-var	1	ϕ (empty set)
1-var	6	{Pkt}, {Trl}, {Pol}, {Gol}, {Siv}, {Dsc}

	Cardina lity	Subsets
2-var	15	{Pol, Pkt}, {Pkt, Trl}, {Pol, Trl}, {Gol, Trl}, {Dsc, Trl}, {Pkt, Gol}, {Pol, Gol}, {Pkt, Siv}, {Trl, Siv}, {Pol, Siv}, {Gol, Siv}, {Dsc, Siv}, {Pkt, Dsc}, {Pol, Dsc}, {Gol, Dsc}
3-var	20	{Pol, Pkt, Dgt}, {Pkt, Gol, Dgt}, {Pol, Gol, Dgt}, {Pkt, Dsc, Dgt}, {Pol, Dsc, Dgt}, {Gol, Dsc, Dgt}, {Pol, Pkt, Gol}, {Pol, Pkt, Siv}, {Pkt, Dgt, Siv}, {Pol, Dgt, Siv}, {Gol, Dgt, Siv}, {Dsc, Dgt, Siv}, {Pkt, Gol, Siv}, {Pol, Gol, Siv}, {Pkt, Dsc, Siv}, {Pol, Dsc, Siv}, {Gol, Dsc, Siv}, {Pol, Pkt, Dsc}, {Pkt, Gol, Dsc}, {Pol, Gol, Dsc},
4-var	15	{Pol, Gol, Pkt, Dgt}, {Pol, Siv, Pkt, Dgt}, {Pol, Dsc, Pkt, Dgt}, {Pkt, Gol, Dsc, Dgt}, {Pol, Gol, Dsc, Dgt}, {Pol, Gol, Pkt, Siv}, {Pol, Dsc, Pkt, Siv}, {Pkt, Gol, Dgt, Siv}, {Pol, Gol, Dgt, Siv}, {Pkt, Dsc, Dgt, Siv}, {Pol, Dsc, Dgt, Siv}, {Pkt, Gol, Dsc, Siv}, {Dgt, Gol, Dsc, Siv}, {Pol, Gol, Dsc, Siv}, {Pol, Gol, Pkt, Dsc},
5-var	6	{Pol, Gol, Pkt, Dsc, Dgt}, {Pol, Gol, Pkt, Dgt, Siv}, {Pol, Dsc, Pkt, Dgt, Siv}, {Pkt, Gol, Dsc, Dgt, Siv}, {Pol, Gol, Dsc, Dgt, Siv}, {Pol, Gol, Pkt, Dsc, Siv},
6-var	1	{Pol, Gol, Pkt, Dgt, Siv, Dsc},

Forward selection reduces the domain space of "*all possible subsets*" approach to at most for consideration as compared to models of the latter [16]. The selection criterion best adapts to large values of p, say 6 in this study, since it is impracticable to traverse all the distinct sub-models which *total* up to 21 occurrences. Besides, Figure 7 gives a pictorial scale for comparing forward/backward selection against "*all possible subsets*".

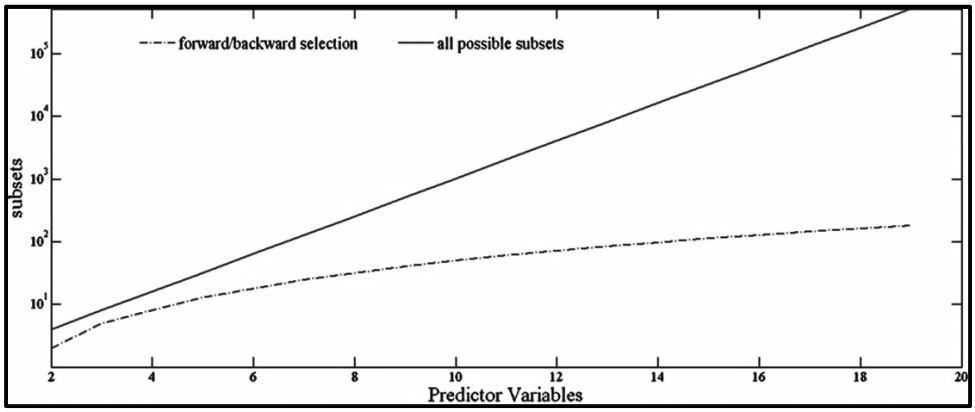

Figure 7. *Comparison of "all possible subsets" with forward selection showing that* $((p^2) / 2) < 2^p$.

Ascending values of *AIC* suggests worse fit but remains more relevant than fewer variables that constrict values of response to a limited number of variables. Similarly, backward selection starts by first regarding the full model as basis for comparison with next lower variable model by dropping one predictor variable. If the succeeding lower variable model has a higher *AIC* value than the full model then focus is shifted and attention is given to the full model thus stopping further considerations. The aim is to find the *lowest AIC* values [16].

On a comparative scale the (*Trailer* = 7) / *Goldbach* sample best model has the *lowest* value of *AIC* and its regression equation is expressed in terms of *Trailer*. Therefore, primes interim formula is given by (25):

$$prime = 5490083 - 1904.54Trl + 416.67Trl^2 - 2.39136Trl^3 \quad (25)$$

The results of these equations for strata of sample spaces (i.e., *Digit*s, *Trailer and Polignac*) as well as the population, as captured by Table 5, and Tables 10-11, are plotted against the generated primes as shown by Figures 8 - 11. The plots exhibit time series trend and seasonality features as can be observed in the *MATLAB* graphs.

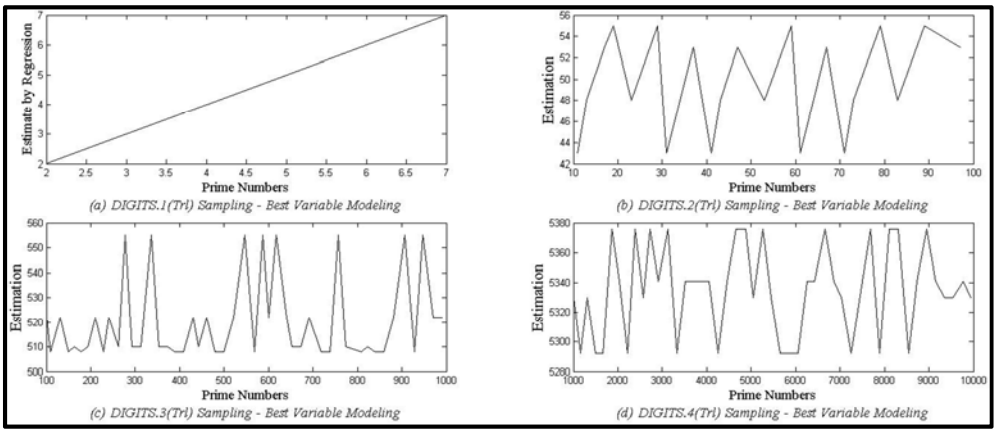

Figure 8. *AIC best model for Digits of values ((1, 4), (2, 21), (3, 143), (4, 1061)).*

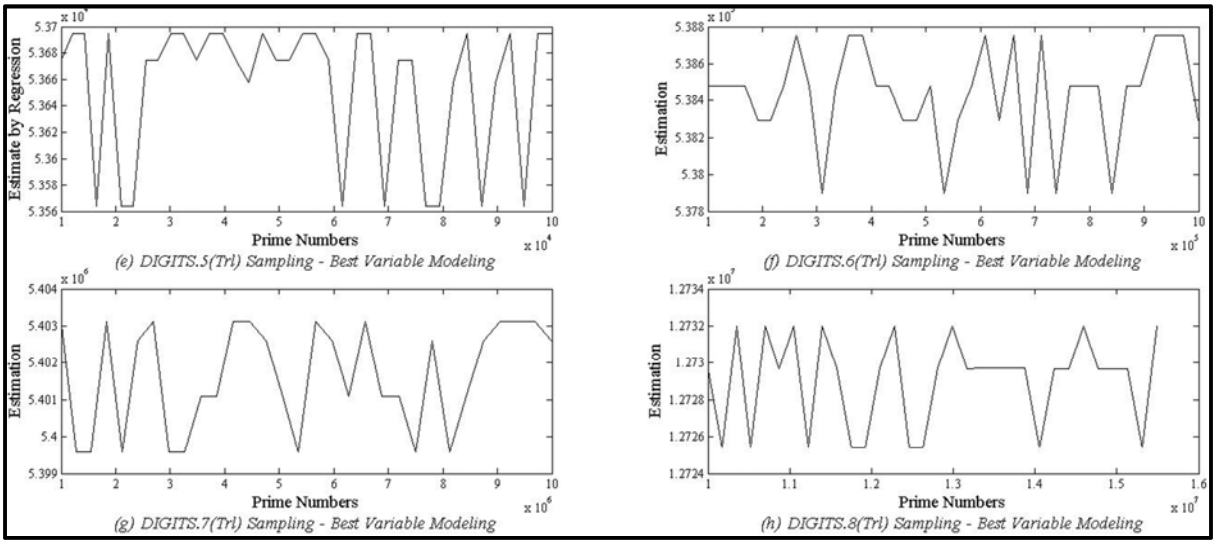

Figure 9. *AIC best model for Digits of values ((5, 8363), (6, 68906), (7, 586081), (8, 335421)).*

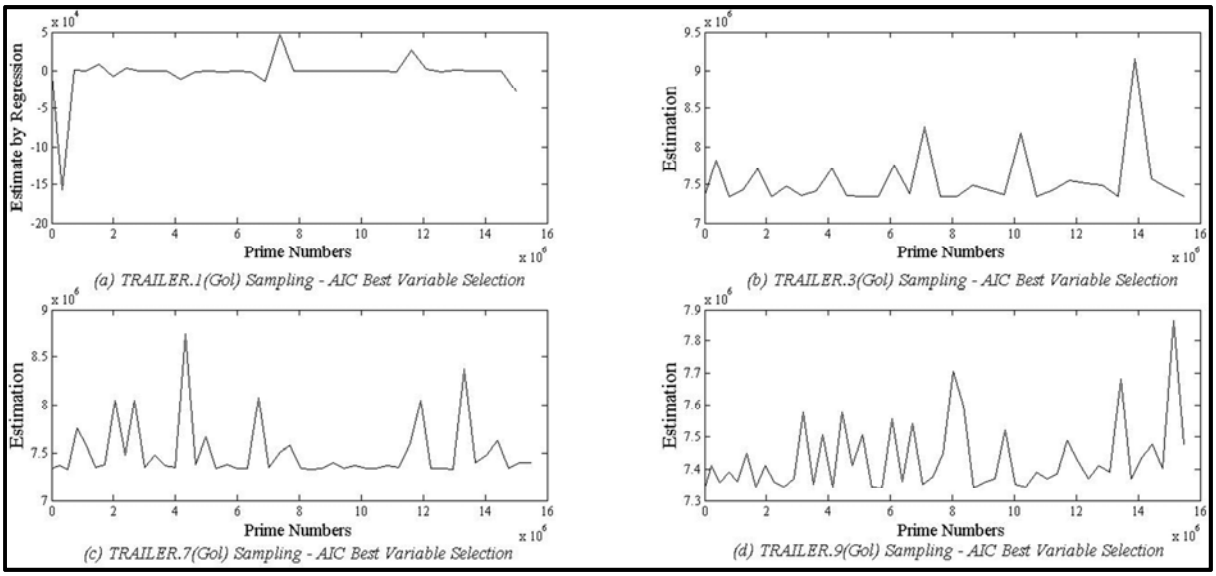

Figure 10. *Best AIC model for frequency and Trailers pairs ((249934, 1), (250110, 3), (250014, 7), (249940, 9)).*

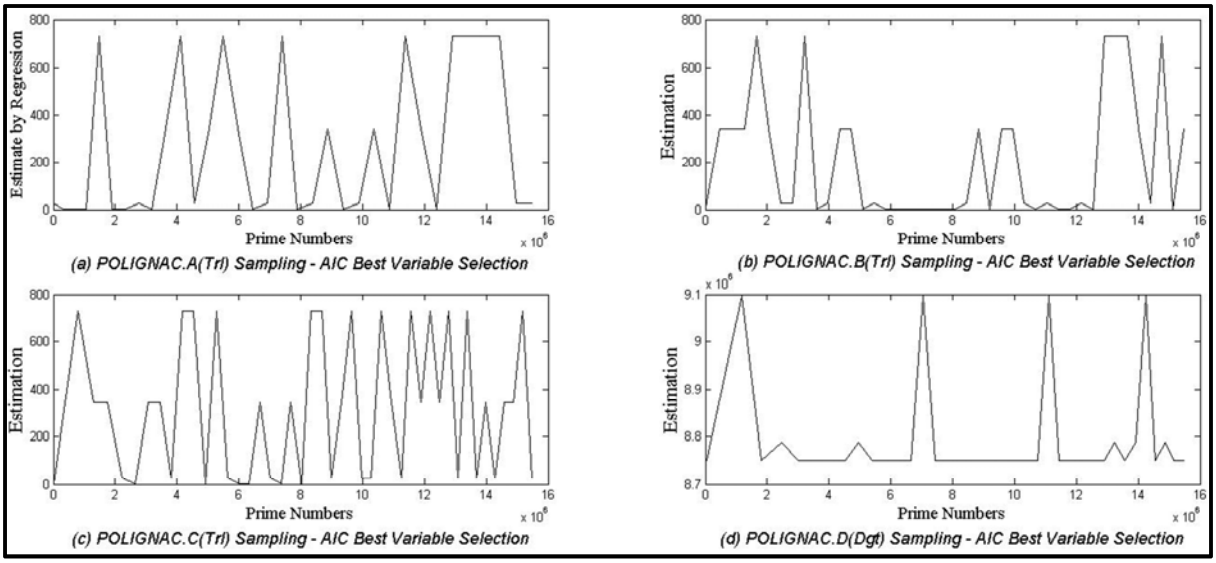

Figure 11. *AIC model for Polignac constants of the range (1-11(A), 12-22(B), 23-33(C), 34-77(D)).*

To gain a better description of the domain space of the entire population, computing of aggregate summaries according to the strata of predictors *Digit*s, *Trailer* and *Gap* (respectively) evolves to capture of Table 5 and Tables 10-11 below. Note: logical *comparisons* = *cmp*, *Goldbach* = *Goldbak*.

Table 10. *Trailer sample summary excluding digits=10 data set.*

Trailer	Total SNo	Lowest cmp	Mean cmp	Total cmp	Highest cmp	Mean Polignac	Total Polignac	Lowest Goldbak	Highest Goldbak	Frequency
1	1645117784193020	13	6804	195145794607	145	10	299376151	-278	278	28679877
3	1645218514404350	7	6805	195171037141	146	10	281864401	-274	288	28681584
7	1645237422689280	9	6804	195158144753	144	11	305518292	-264	256	28681703
9	1645200562420740	15	6805	195167586325	160	10	290627479	-288	272	28680626

Table 11. *Polignac gap sample summary according to class interval of size 16 as expressed by (5).*

Polignac Interval	Lowest SNo	Mean SNo	Lowest Prm	Lowest Pkt	Total cmp	Mean cmp	[b] Lowest Goldbak	[b] Highest Goldbak	Frequency
1 - 16	2	56704661	3	7	631559864636	6755	-30	288	93500005
17 - 32	218	59690421	1361	33	123558163319	6983	-62	254	17693875
33 - 48	3386	62640834	31469	90	21673336050	7199	-94	240	3010664
49 - 64	31546	65309196	370373	232	3238585234	7386	-126	110	438453
65 - 80	104072	67810512	1357333	393	518076388	7558	-158	44	68550
81 - 96	1094422	70034620	17051890	1145	80824282	7710	-190	2	10483
97 - 112	1319946	70934072	20831530	1245	11846691	7773	-222	-68	1524
113 - 128	10539433	69817893	189695900	3267	1546683	7733	-250	-92	200
129 - 144	23163299	85021168	436273300	4707	283479	8590	-276	-206	33
145 - 160	72507381	96030208	1453168000	8060	36685	9171	-288	-260	4

[b]lowest Goldbak=min(Goldbach), highest Goldbach=max(Goldbach)

As can be corroborated by Figure 12, if a listing of 67 consecutive primes (by sequence numbers) is made, the probability of having instances where *Cycle* = 2 is higher than the case of *Cycle* = 3. The rarest instances in this listing occurs where *Cycle* = 1. In this set, at least twice as much instances with *Cycle* = 2 is expected. On the other hand, considering the sample with focus on last two *digit*s of a *prime* pair (*Trailer2*), Figure 12 indicates that the composites are the rarest (39 for composites) whilst the primes are the most spread.

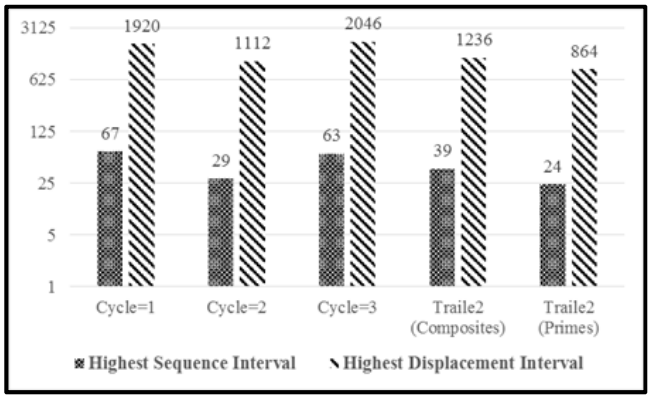

Figure 12. *Maximal Consecutive Primes sequence and distance within which a member of these samples must be encountered.*

With respect to displacement by value from an orderly listing between consecutive neighboring pairs of twin primes, Figure 11 shows the distribution amongst three groups of non-intersecting data sets with respect to label of standalone workstation at one's disposal for parallel processing. The peak values of the respective *gap*s are displayed on the graph with *Cycle*=3 having the *highest* possible value.

As the results indicate (s_ALL_*prm*), 514 is the true value Zhang assumed as the hypothetical 7 10^7 [30]. As for the *Trailer2* parameter, it is instructive to specify that the class of composites are at most 1236 apart whilst 864 is for the class of primes as seen in Figure 12.

6.2. Hypothesis Testing

Perfection of prediction is practically unrealizable as it has an attendant error cost as measured by standard error of estimate. Assuming 0.01 is the significance level for this empirical study with a very high requirement for accuracy in the observations - this necessitates that $(1 - R^2 < 0.01)$. For optimization of prediction, population data is subjected to maximization of "*(adjusted) R squared*" with a corresponding minimization of standard error whilst ensuring that *ρ-values* of minimal variables are less than significance level of 0.01.

These criteria are best satisfied by the subsets depicted in Table 12.

Table 12. *Candidates for prediction – cubic polynomials.*

Subsets	Standard Error	R Square	SSE
Dsc, Siv	1355.04	0.99999983	8.2947852759E+18
Pkt, Dsc, Trl	1597.53	0.9999998	8.2947847390E+18
Pkt, Gol, Dsc, Trl	1597.39	0.9999998	8.2947847393E+18
Pol, Dsc, Pkt, Trl	1596.61	0.9999998	8.2947847412E+18

Recall that cyclotomic polynomial of degree 3 does not exist. As a consequence, the quartic functions provide the next upper level for formulating a better solution to this prediction

problem. Degree 4 cyclotomic polynomials are given by: $\varphi_5(x) = x^4 + x^3 + x^2 + x + 1$, $\varphi_8(x) = x^4 + 1$, and $\varphi_{10}(x) = x^4 - x^3 + x^2 - x + 1$. To be consistent with this direction of progression, the powers of the predictor variables in Table 12 are extended to the fourth index to effectively realize and maintain this deviation. This expansion has the advantage of maximizing the utilization of the capacity of 16 variables provided by Excel without any consequential conflict of storage limit. Transformation into polynomials of degree 4 requires the elimination of variables from Table 14 that did not yield *ρ-values* less than the significance level of 0.01 (e.g., *Trailer*), maintaining those that conform to optimization requirements and introducing new ones (e.g., *Polignac* replacing *Gol* in {*Pkt, Gol, Dsc, Trl*}. Hence, transition from Table 12 into Table 13 results, with noticeable improvements in standard error, *R square* and *SSE* values (26).

Table 13. *Candidates for prediction – quartic powers of polynomials.*

Subsets	d[a]	Df[b]	Standard Error	R Square	SSE
Dsc, Siv	3	6	1355.039906	0.999999833982	8.29478527597E+18
Pkt, Dsc		6	1686.579603	0.999999742803	8.29478451967E+18
Pol, Pkt, Dsc		9	1686.042872	0.999999742968	8.29478452103E+18
Pol, Gol, Pkt, Dsc		12	1686.027407	0.999999742974	8.29478452108E+18
Dsc, Siv	4	8	1353.848194	0.999999834274	8.29478527840E+18
Pkt, Dsc		8	1686.400917	0.999999742858	8.29478452012E+18
Pol, Pkt, Dsc		12	1684.947428	0.999999743303	8.29478452381E+18
Pol, Gol, Pkt, Dsc [b]		14	1684.930617	0.999999743309	8.29478452386E+18

[*a] d = maximum power index of var terms considered (degree of polynomial), df = degree of freedom (no of terms of predictor var(s)).
[b] df = 14 (Gol^2 and Gol^4 terms are dropped since their ρ-values are greater than 0.01, hence are statistically insignificant- accounting for why df≠16).

$$\begin{aligned} \Pr imes = {} & 1434.9686 + 39.286491\text{Dsc} + 1.1885535 Dsc^2 - \\ & 9.913634 Dsc^3 + 6.624045 E\text{-}10 Dsc^4 - 1.958679\text{E}2\text{Siv} + \\ & 6.0718767\text{Siv}^2 + 1.963455\text{E}\text{-}3\text{Siv}^3 - 3.190334\text{E}\text{-}7\text{Siv}^4 \end{aligned} \quad (26)$$

Amongst the four candidates vying for selection as prime's formula, the subset {*Dsc*, *Siv*} satisfies the criteria for minimal of both standard error and AIC value as well as maximal of both SSE and PVE values. Table 14 shows *ρ-values* satisfying conformity with significance level less than 0.01.

Table 14. *Significance levels of {dsc, siv} subset.*

	S.N	Coefficients	ρ-value
Intercept	1	1434.9686	6.24E-123
Dsc	2	39.286491	8.85E-299
Dsc^2	3	1.1885535	0
Dsc^3	4	-9.914E-06	0
Dsc^4	5	6.624E-10	6.49E-93
Siv	6	-195.86797	0
Siv^2	7	6.0718767	0
Siv^3	8	0.0019635	2E-153
Siv^4	9	-3.19E-07	0.0003027

Statistically, in order to associate predictors to a response variable in (9) and (24), the initial assumption is that there is no basis of using the independent variables to predict a given response variable through null hypothesis testing:

H_0: $\beta_1 = 0$ & $\beta_2 = 0$ & $\beta_3 = 0$ (Population correlation = 0) vs.

H_1: either $\beta_1 \approx 1435 \neq 0$ or $\beta_2 \approx -39.2865 \neq 0$ or $\beta_3 \approx 1.1885535 \neq 0$, etc. (correlation coefficients not zero indicates that predictors are valid for estimation of response variables, called alternative hypothesis).

The null hypothesis is tested by the condition "all $\beta_i = 0$" against the hypothesis "at least a $\beta_i \neq 0$"

Let us assume significance level of 0.01. A lower *ρ-value* than 0.01 clearly indicates that rejection of the null hypothesis in favor of the alternative hypothesis. Normally, investigation for a hypothesized relationship demands that the *ρ-value* of 0.01 or 0.05 is initially set in advance of the empirical research study [10, 18]. Table 14 shows that results from the study all yield *ρ-value* less than that specified in advance.

Therefore, this study is significant in that it concludes that a relationship really exists for associating the variables. Figure 13 is a visual representation for the estimates of (24) against observed primes.

6.3. Interpretation of Figures

It is quite trivial, for the population space, to discuss how many primes are involved because of the sequential nature of the listing, i.e., 1. However, for other sample spaces, Figure 11 gives the corresponding values in terms *SNo* displacement obtained from tracking of population data.

For the population data (ALL), it is clear that the distance (value-wise) between successive *prime* pairs cannot exceed 514 in all cases. The value 2046 appears to be the *highest* distance possible for the sample spaces under consideration (as is depicted in Figure 12).

Furthermore, in order to gain a better description of the domain space of the entire population, aggregate summaries according to the strata of the predictors must be computed: *Digit*s, *Polignac*, *Goldbach* and *Trailer* as captured by Tables 5, 10-11. However, both Figure 8 and Figure 13 give similar predictions with fault tolerable smoothing.

Observation shows that a line of fit in Figure 13 plot was determined by MATLAB which perfectly smoothens these seasonal fluctuations. Displaying plots with lines of fits for the other samples (e.g., Figure 13 for the sample Dgt11U) will be monotonous representations and so is not shown.

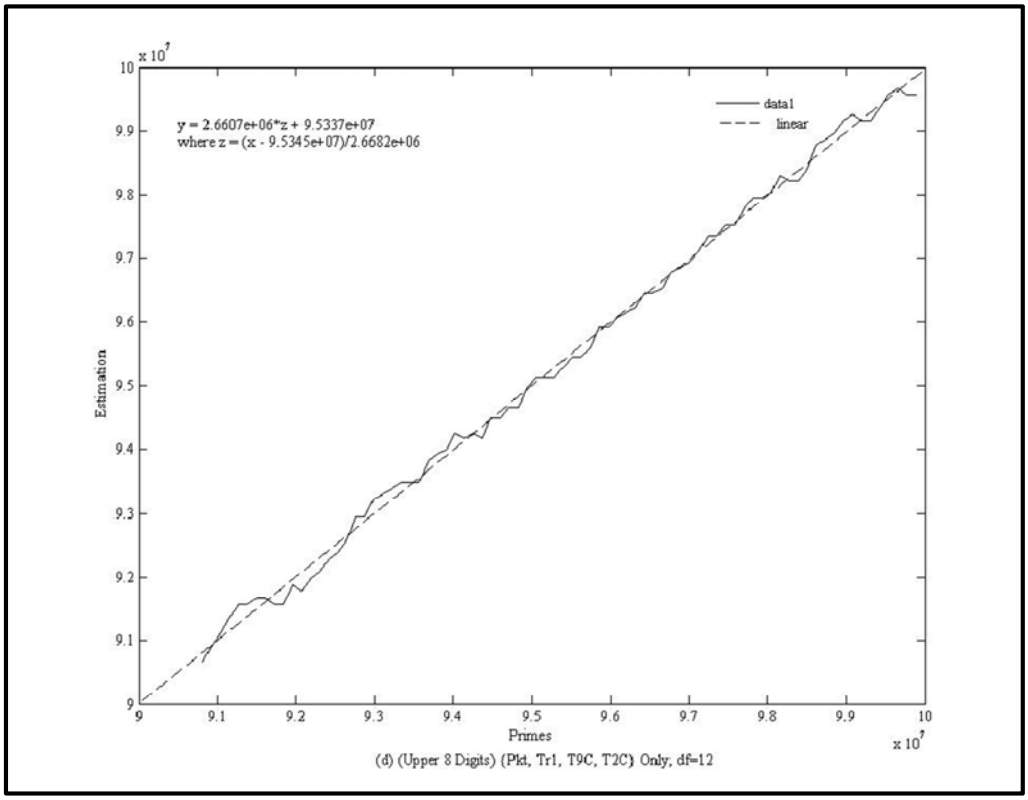

Figure 13. *AIC Best modeling (variable selection) for upper sample records where Digits=8 and features selected (pkt, trl, t9c) through Best Regression Formula with DF = 12 and fitted to the line y = 2.6607e+06*z + 9.5351e+07 where z = (x – 9.53452e+07) / 2.6682e+06.*

Figure 14 depicts monotonically increasing time complexity with corresponding digits increments. This suggests the need for increasing logical comparison operations with growth in digits.

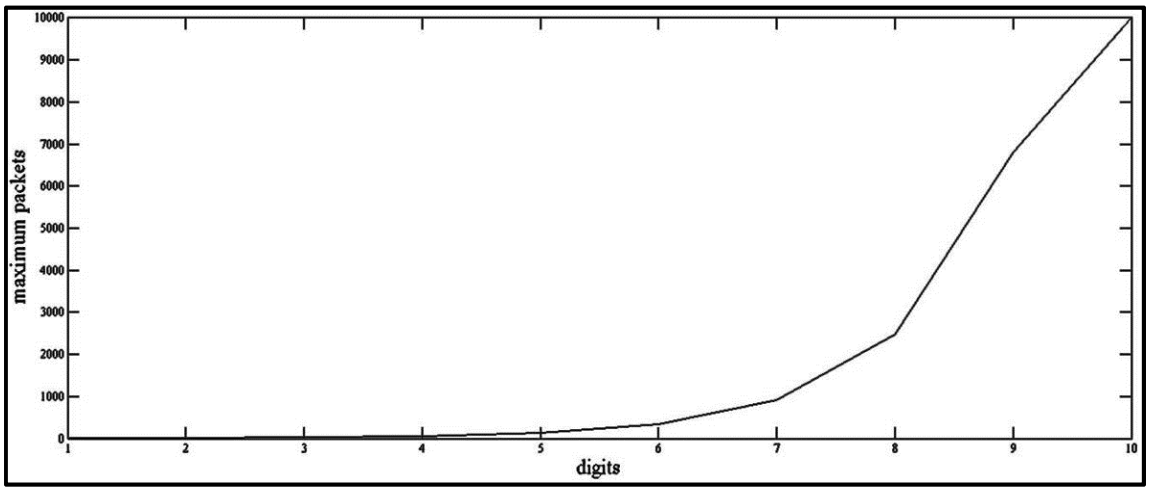

Figure 14. *Graph showing direct proportional relationship between number of digits and worst case time complexity.*

AIC criterion for variables selection built best models for respective sample spaces but on comparison with actual *prime* values there exist a lot of *gap*s in the prediction for the entire domain space as evidenced by fluctuations of the curves. This suggests that no sample accurately represents the entire *prime* behavior but just a segment. Figures 8 –13 captured the various seasonal trends, of order (lag) k - correlational dependency of an i-th element of a time series with its successors of width in multiples of k. Seasonality is identified by repetition of a curve pattern with a minimal error in measurements [21].

The results of this research confirm that *Trailer* is the most significant variable for testing primality although outstanding research works did not adequately address this requirement (Figure 15 shows the equity and balanced spread of peak *gap*s). In descending order of significance, *gap*s of *Polignac* and *Goldbach* follow next whilst *Packets*, *Discarded* and *Siv* are positioned as third, fourth and fifth respectively.

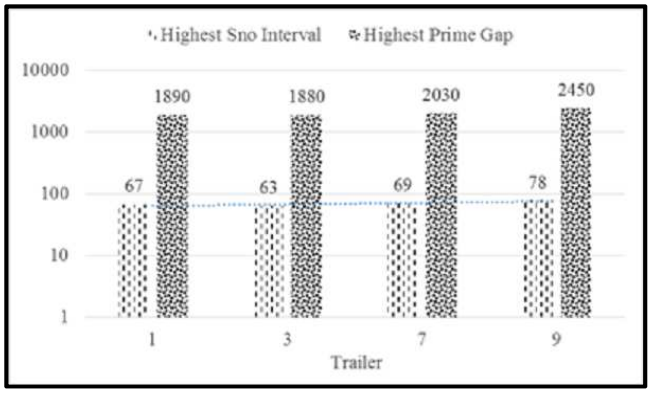

Figure 15. *Distribution of Gaps with respect to SNo and distances between prime pairs.*

On a very critical appraisal, AIC modeling failed to take into account that the parameter significance F which on maximization, can substitute the AIC formula and achieve the same objective in best model construction. In fact, the maximal value of significance F is equivalent to the minimal AIC value being sought after. It can be argued that this approach is a duplication of an earlier effort. Moreover, forward/backward selection in some occasions may lead to errors in modeling. The safest approach is the "all possible subset" which is exhaustive in its search techniques.

6.4. Tables Analyses

Table 5 shows that the higher the *digit*s the more complex the algorithm becomes in time bounds of "*lowest comparison*", "*highest comparison*", and "*total comparison*" and the more the frequency count of primes generated as supported by "*lowest SNo*", "*highest SNo*" and "*total Prime*" (Figure 15 depicts the spread of the data). Others with the same trend are "*highest Polignac*", "*total Polignac*", "*lowest Goldbach*", "*highest Goldbach*", and "*total Prime*".

"*Polignac Interval*" is distributed for the range (1-160) with a sub-unit at every 16 value. In Table 10, data for the aggregates "*lowest SNo*", *"mean SNo*", "*lowest prime*", "*lowest comparison*" and "*mean comparison*", are monotonically increasing with corresponding progression in "*Polignac Interval*" except in the ranges 97-112 and 113-128 where there are abnormal variations (outliers) due to noises (disturbances) that hinder us from producing a smooth curve. Meaning that as both lower and upper limits of "*Polignac Interval*" increase, a commensurate increase in the values of the aggregates mentioned earlier is experienced. A direct proportion relationship exists for the aggregates with the "*Polignac Interval*".

On the other hand, monotonically decreasing curve is observed in the summaries "*total comparison*", "*lowest Goldbach*", "*highest Goldbach*", Frequency - increase in lower and upper limits of "*Polignac Interval*" yields lower values of the summaries just outlined. However, inverse is the case for ("*total comparison*", "*lowest Goldbach*", "*highest Goldbach*", *Frequency*) parameters (depicted by Table 11). Figure 16 captures the spread of prime's records across *Polignac* coefficients. Closely observing the graph reveals that peak values were recorded consistently at 3, 6, 9, 12 and 15: thus suggesting that applying modulo3 arithmetic on the coefficient with military-like accuracy is much desired. An implication of these findings is that the parameters (*G_Pm1_qqq, G_Pm2_qqq, G_Pm3_qqq), (P_Bm1_qqq, P_Bm2_qqq, P_Bm3_qqq), (P_Gm1_qqq, P_Gm2_qqq, P_Gm3_qqq*) have adequate justification for existence and consideration for further analysis, where qqq stands for *SNo* and *Prm*, *P* for *Polignac* and *G* for *Goldbach* (refer to Table 1 for details).

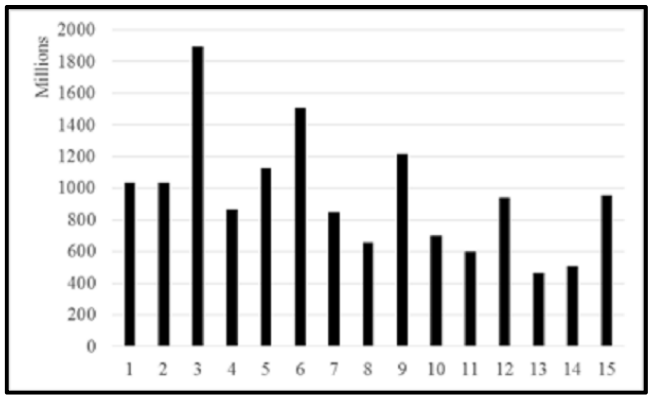

Figure 16. *Spread of primes across lower 15 Polignac coefficients.*

For this algorithm analysis, what constitutes important cost of operation for assessing efficiency is comparison operations as associated with control statements (e.g., while, repeat, for, if etc.). An Execution instance of any of these statements is counted as 1 Packet. Formally, algorithm efficiency is a measure of time requirement in terms of problem size expressed as in terms of constituent elements in the application. If there exists a threshold value, serving as an upper bound time, upon which processing of an n-size problem can never be exceeded by an algorithm, then the worst case limit is coined(Θ) - behavioral analysis hereby adopted for making decisions about algorithm performance. There is a window for assessing properties of estimators with the presumption of indefinite growth of sample size as n tends towards infinity. Asymptotic results are as valid as those of finitely sized samples [14].

Distributions of "*mean comparison*", "*total comparison*", "*mean Polignac*", "*lowest Goldbach*" & Frequency is fairly and equitably spread - the differences amongst pair is not much significant to be alarmingly and considerably noted. However, for the parameters "*lowest comparison*", "*highest Polignac*" and "*highest Goldbach*", maximal values were recorded at *Trailer* = 9. All these associations is tracked at Table 11.

From Table 11, it is clear that least minimal "*highest Polignac*" (144) and least minimal "*highest Goldbach*" (256) is only found where *Trailer*=7 and no other. Cryptanalysts are bound to take advantage of this opportunity as there is no worst case time complexity but the best case here. Therefore, primes under this category are most likely to be easily divisible in a very short span of execution time. Also, *Trailer*=7 has the *highest* value (maximal) of "mean

Polignac". For the record where *Trailer*=9, most maximal "*highest Polignac*" (160) and least minimal "*lowest Goldbach*" (-288) are traced here – implying that worst case and hardest divisibility operations are located here. Cryptologists are most likely to choose this as "very hard" primes.

Any of the four regression formula constructed from subsets of predictors in Table 14 is capable of predicting all possible primes although with tolerable errors of computation. These formulae established that relationship truly exists among these pairs of subsets of predictor variables with prime numbers (response variable) since the alternative hypothesis had been proved as statistically significant in the four cases (although (25) is preferred). The *prime* generating algorithm is quite revealing unlike the views of Euler since there exists strong evidence of orderly sequence of patterns guiding the behavior of primes away from the law of chance [11, 19].

7. Conclusion

Unlike Euler's formula with polynomial of degree 2 in terms of just one variable, this study expanded the domain of spread to polynomials of degree 3 in terms of 6 predictor variables in the formulation of prime numbers. Future research should concentrate on quartic functions since cyclotomic functions of degree 4 exist here. However, there will arise restriction on number variables to consider for manipulation. In essence, $4x4 < 3x6$ variables for variable selection was adopted.

Robust and 'hard' primes, that are guaranteed to frustrate efforts of cryptanalysts, is advocated through implementation of the conditions (*Trailer*=9), (*Polignac*>=160), (*Goldbach*<=-288) and (very high decimal *digit*s e.g., 100). This way, the primes are safe from cryptanalysts' nefarious acts and integrity of these computations will be preserved. Relatively 'easy' primes are found where (*Trailer*=7) and (*Polignac*<=144) and (*Goldbach*>=-264) and (*Goldbach*<=256). Cryptologists beware of this best case time complexity scenario - these primes are relatively easily divisible in a very short span of execution time.

AIC model building has its limitations and potential weaknesses. Firstly, its emergence would have been more admired without its failure to accommodate a more conventional statistical measure instead of its newly introduced formula. Significance F is an alternative path to variable selection which AIC formula ignored completely. The subset with maximal value of significance F is equivalent to minimal AIC value for building best model with regards to data fitting. It is only "all possible subsets" that can avoid this error of abnormal variations in subsets. Other restrictions and limitations of AIC modeling include export from MS Access to MS Excel must also not exceed 65,536 (216) records whilst a worksheet cannot exceed 1,048,576 rows and 16 variables are permissible for all possible subsets. Rounding off operation results in loss of precision which primes will signal as intolerable errors. Retracing and re-computing of built-in functions likely to yield truncated results is worth the effort.

Although recursive functions degrade efficiency, however, their power of strong formulation cannot be ignored. Besides, other limitations (e.g., capacity and performance of spreadsheet) need to be acknowledged with speed of processing being the most significant factor.

Equation (2) becomes (27) thus:

$$\lim_{n\to\infty} \inf(p_{n+1} - p_n) < 514 \tag{27}$$

Twin primes conjecture now has a definite bound within which it is true without wild goose chase. However, generalizability of this finding is constrained within the first 20 billion lower primes. As can be seen from Figure 17 below, the monotonically increasing trend confirms that the value of 514 may not hold for *digit*s higher than 12. Outside this range, the generalizability of this finding cannot be guaranteed but confined to the first 20 billion lower primes considered.

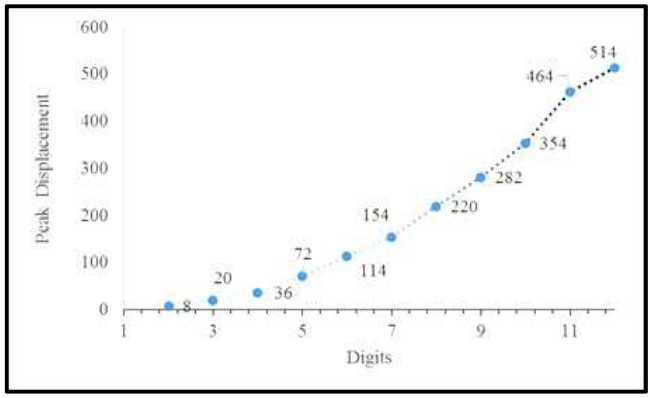

Figure 17. *Tracking of peak distances amongst prime pairs spread across digits.*

This study recognized and supported outstanding works on the significance of number of *digit*s to adding complexity to prime numbers but extend this further to the most significant *digit*. Further research should be done to investigate the fragmentation of the *Lead* parameter with a view to exploring its behavioral patterns that govern primes.

Data eccentricity in confronting the prime numbers computational problem heavily guided this work study.

Acknowledgements

This research was carried out in affiliation with Universiti Teknologi Petronas (Malaysia) and appreciations and gratitude is hereby extended to all parties. Moreover, regards is also extended to Ibrahim Badamasi Babangida University, Lapai (IBBUL) – Nigeria, for sponsoring me under TETFUND/ES/AST&D/IBBU/LAPAI/VOL. I dated Sep, 28 2012. My sincere gratitude goes to all that contributed to this work and they are too numerous to mention.

References

[1] L. M. Adleman and M. A. Huang, "Primality testing and abelian varieties over finite fields," *Lecture Notes in Mathematics*, vol. 1512, Springer-Verlag, 1992.

[2] L. M. Adleman, C Pomerance and R. S. Rumely, “On distinguishing prime numbers from composite numbers,” *Annals of Mathematics*, vol. 117, pp. 173-206, 1983.

[3] M. Agrawal, N. Kayal and N. Saxena, "Primes is in P," *Annals of Mathematics,* vol. 160 (2), pp. 781-793, 2004.

[4] W. S. Anglin, "The square pyramid puzzle," *The American Mathematical Monthly*, Mathematical Association of America, vol. 97 (2), pp. 120-124, 1990. doi: 10.2307/2323911, http://www.jstor.org/stable/2323911.

[5] A. L, Frank, "An overview of elliptic curve primality proving," *In Other Words 2*, 2011. Available online: http://www.stanford.edu/class/cs259c/finalpapers/primalityproving.pdf.

[6] W. Bosma and M. van der Hulst, “Primality proving with cyclotomy,” PhD thesis: Universiteit van Amsterdam, 1990.

[7] T. F. Boucher, "On cyclotomic primality tests," Master’s thesis, University of Tennessee, 2011.

[8] D. M Bressoud, “Factorization and primality testing,” Springer Science and Business Media, 2012.

[9] G. A. Einicke, “Smoothing, filtering and prediction: estimating the past, present and future,” Intech: Janeza Trdine, 2012.

[10] D. A. Freedman, Statistical Models: Theory and Practice, Cambridge University Press, 2009.

[11] J. Havil, “Gamma: exploring euler's constant,” *The Mathematical Intelligence*, Princeton University Press, pp 1-260, 2009.

[12] B. Gates, The Road Ahead. New York: Viking, 1995.

[13] S. Goldwasser and J. Kilian, "Almost all primes can be quickly certified," in *Proc. 18th STOC*, pp. 316-329, 1986.

[14] R. Busatto, “Using time series to assess data quality intelecommunications data warehouses,” in *Proceedings of the 2000 Conference on Information Quality*, pp. 129-136, 2000.

[15] Solovay, R. M. Strassen and Volker, “A fast Monte-Carlo test for primality,” SIAM Journal on Computing,” vol. 6 (1), pp. 84-85, 1977 doi: 10.1137/0207009/ Available online: http://www.revolvy.com/topic/Solovay-Strassen%20primality%20test&item_type=topic.

[16] L. L Nathans, F. L. Oswald and K. Nimon, “Interpreting multiplelinear regression: a guidebook of variable importance,” *Practical Assessment, Researchand Evaluation*, vol. 17 (9), 2012. Available online: http://pareonline.net/getvn.asp?v=17&n=9

[17] J. Jakun, Registered Certificates of Primality, 1975. Available online: wwwmath.anu.edu.au/~brent/pd/AdvCom2t.pdf Posted: http://jeremykun.com/2013/06/16/miller-rabin-primality-test/

[18] E. W. Weisstein, “Prime number,” *MathWorld*, 2004. Available online: http://mathworld.wolfram.com/PrimeNumber.html.

[19] J. Hoffstein, J. Pipher and J. H. Silverman, An Introduction to Mathematical Cryptography, Springer Science and Business Media, 2nd ed, 2014.

[20] Y. Motohashi, The Twin Prime Conjecture, Marh. NT, 2014. Available online: www.math.cst.nihon-u.ac.jp/~ymoto/.

[21] D. F. Andrews, “A robust methodfor multiple linear regression,” *Journal of Technometrics* (Amrtican Statistical Association), vol. 16 (4), pp. 523531, 2012. Available online: http://dx.doi.org/10.1080/00401706.1974.10489233.

[22] http://www.theoremoftheday.org/LogicAndComputerScience/Pratt/TotDPratt.pdf, accessed on: 201403241130.

[23] D. H. Lehmer, “An extended theory of Lucas’ functions,” *Annals of Mathematics*, vol. 31 (3), pp. 419–448, 1930.

[24] L. C, Washinton, Elliptic Curves: Number Theory and Cryptography. 2nd ed, Chapman and Hakk/VRC, 2008.

[25] D. A. Lind, W. G. Marchal and S. A. Wathen, Statistical Techniques in Business & Economics, McGraw-Hill, NY: Irwin, pp. 461-542, 2012.

[26] G. L. Miller, “Riemann’s hypothesis and tests for primality,” *Journal of Computer and System Sciences*, vol. 13 (3), 1976.

[27] K. H. Rosen and K. Krithivasan, Discrete Mathematics and Its Applications, 7 ed., McGraw-Hill, NY: Connect Learn Succeed, pp. 237-306, 2013.

[28] R. Schoof, "Four primality testing algorithms," *Algorithmic Number Theory, MSRI Publications*, vol. 44, pp. 101-126, 2008.

[29] R. Solovay and V. Strassen, "A fast Monte-Carlo test for primality," *SIAM J. Comput*. vol. 6。

[30] Y. Zhang, "Bounded Gaps Between Primes," *Annals of Mathematics*, JSTOR Publications, vol. 179 (3), pp. 1121-1174, 2014.

Biography

Bashir Kagara Yusuf completed both B.Tech and M.Sc (in Computer Science) honors degrees from Abubakar Tafawa Balewa University (ATBU), Bauchi - Nigeria, in the years 1994 and 1997 respectively. Nigerian Telecommunications (NITEL) then employed him as System Analyst/Programmer rising to the position of Assistant Manager in the Information Technology department. He also rose to the rank of Assistant Lecturer in the Mathematical Sciences (ATBU) in 1998 and is actuvely engaged in teaching undergraduate computer science courses. Administering Sybase and Oracle databases remains his specialty with strong bias in Perl programing amongst others. Currently at Ibrahim Badamasi Babangida University, Lapai-Nigeria.

Ahmad Kamil Bin Mahmood has over twenty years of experience in various fields of Information Systems. He obtained his B.Sc. (Hons) in Actuarial Science from University of Iowa in 1986, MSc in Actuarial Science and Statistics from the same university in 1988, and PhD from Salford University, Manchester, UK in 2005. He currently holds the Head of High Performance Cloud Computing Centre (HPC3) at the Universiti Teknologi PETRONAS (UTP). His current research areas are in cloud computing, e-services quality and measurement, artificial intelligence application, and knowledge management. He has published over 100 papers in these areas.